

08/836734

95104575

PCT/EP



Europäisches  
Patentamt

European  
Patent Office

Office européen  
des brevets

RECD 29 JAN 1996  
WIPO PCT

Bescheinigung      Certificate      Attestation

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr.    Patent application No.    Demande de brevet n°

94402668. 1

PRIORITY DOCUMENT

Der Präsident des Europäischen Patentamts;  
Im Auftrag

For the President of the European Patent Office  
Le Président de l'Office européen des brevets  
p.o.

H. B. m.

H. Raaphorst

Den Haag, den  
The Hague,  
La Haye, le 15/01/96



# **Europäisches Patentamt**

**European  
Patent Office**

## **Office européen des brevets**

**Blatt 2 der Bescheinigung  
Sheet 2 of the certificate  
Page 2 de l'attestation**

**Anmeldung Nr.: Application no.: Demande n°:** 94402668.1

Anmeldetag:  
Date of filing:  
Date de dépôt: 22/11/94

Anmelder:  
Applicant(s):  
Demandeur(s):  
**ASSOCIATION FRANCAISE CONTRE LES MYOPATHIES**  
**F-75013 Paris**  
**FRANCE**

**Bezeichnung der Erfindung:**  
**Title of the invention:**  
**Titre de l'invention:**  
**LGMD gene**

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s) revendiquée(s)

**Staat:** **Tag:** **Aktenzeichen:**  
**State:** **Date:** **File no.**  
**Pays:** **Date:** **Numéro de dépôt:**

**Internationale Patentklassifikation:**  
**International Patent classification:**  
**Classification internationale des brevets:**  
**C12N15/57, C12N9/64, A61K38/48, G01N33/50, C12Q1/68, C12N5/10, A61K48/00**

**Am Anmeldestag benannte Vertragstaaten:**  
Contracting states designated at date of filing: AT/BE/CH/DE/DK/ES/FR/GB/GR/I/E/IT/LI/LU/MC/NL/PT/SE  
**Etsats contractants désignés lors du dépôt:**

**Bemerkungen:**  
**Remarks:**  
**Remarques:**

1  
**LGMD gene**

4

The invention relates to the isolated gene coding for a calcium dependent protease belonging to the Calpaïn family which, when it is mutated, is a cause of 5 a disease called Limb-Girdle Muscular Dystrophy (LGMD).

The term limb-girdle muscular dystrophy (LGMD) was first proposed by Walton and Nattrass (1954) as part of a classification of muscular dystrophies. LGMD is characterised by progressive symmetrical atrophy and weakness of the proximal limb muscles and by elevated serum creatine kinase. Muscle biopsies 10 demonstrate dystrophic lesions and electromyograms show myopathic features. The symptoms usually begin during the first two decades of life and the disease gradually worsens, often resulting in loss of walking ability 10 or 20 years after onset (Bushby, 1994). Yet, the precise nosological definition of LGMD still remains unclear. Consequently, various neuromuscular diseases such as 15 facioscapulohumeral, Becker muscular dystrophies and especially spinal muscular atrophies have been occasionally classified under this diagnosis. For example, a recent study (Arikawa et al., 1991) reported that 17% (out of 41) of LGMD patients showed a dystrophinopathy. These issues highlight the difficulty in undertaking an analysis of the molecular and genetic defect(s) involved in this 20 pathology.

Attempts to identify the genetic basis of this disease go back over 35 years. Morton and Chung (1959) estimated that "the frequency of heterozygous carrier ... is 16 per thousand persons". The same authors also stated that "the segregation analysis gives no evidence on whether these genes in different 25 families are allelic or at different loci". Both autosomal dominant and recessive transmission have been reported, the latter being more common with an estimated prevalence of  $10^{-5}$  (Emery, 1991). The localisation of a gene for a recessive form on chromosome 15 (LGMD2A, MIM 253600; Beckmann et al., 1991) provided the definitive proof that LGMD is a specific genetic entity. 30 Subsequent genetic analyses confirmed this chromosome 15 localisation (Young et al., 1992; Passos-Bueno et al., 1993), the latter group demonstrating genetic heterogeneity of this disease. Although a recent study localised a second mutant

gene to chromosome 2 (LGMD2B, MIM 253601; Bashir et al., 1994), there is evidence that at least one other locus can be involved.

Genetic analyses of the LGMD2 kindreds revealed unexpected findings. First genetic heterogeneity was demonstrated in the highly inbred Indiana Amish 5 community. Second although the Isle of la Réunion families were thought to represent a genetic isolate, at least 6 different disease haplotypes were observed , providing evidence against the hypothesis of a single founder effect (Beckmann et al., 1991) in this inbred population.

The nonspecific nosological definition, the relatively low prevalence and 10 genetic heterogeneity of this disorder limit the number of families which can be used to restrict the genetic boundaries of the LGMD2A interval. Cytogenetic abnormalities, which could have helped to focus on a particular region, have not been reported. Immunogenetic studies of dystrophin-associated proteins (Matsumura et al., 1993) and cytoskeletal or extracellular matrix proteins such as 15 a merosin (Tomé et al., 1994) failed to demonstrate any deficiency. In addition, there is no known specific physiological feature or animal model that could help to identify a candidate gene. Thus, there is no alternative to a positional cloning strategy.

It is established that the LGMD2 chromosomal region is localized on 20 chromosome 15 as 15q15.1 - 15q21.1 region ( Fougerousse et al., 1994).

Construction and analysis of a 10-12 Mb YAC contig (Fougerousse et al., 1994) permitted the mapping of 33 polymorphic markers within this interval and to further narrow the LGMD2A region to between D15S514 and D15S222. Furthermore, extensive analysis of linkage disequilibrium suggested a likely 25 position for the gene in the proximal part of the contig.

The invention results from the construction of a partial cosmid map and the screening by cDNA selection (Lovett et al., 1991 ; Tagle et al., 1993) for muscle-expressed sequences encoded by this interval led to the identification of a number of potential candidate genes. One of these, previously cloned by 30 Sorimachi et al. (1989), encodes a muscle specific protein, nCL1 (novel Calpain Large subunit 1), which belongs to the calpain family (CANP, calcium-activated neutral protease; EC 3.4.22.17), and appeared to be a functional candidate gene for this disease.

Calpains are non-lysosomal intracellular cysteine proteases which require calcium for their catalytic activities (for a review see Croall D.E. et al, 1991). The mammalian calpains include two ubiquitous proteins CANP1 and CANP2 as well as tissue-specific proteins. In addition to the muscle specific nCL1, stomach specific nCL2 and nCL2' proteins have also been described; these are derived from the same gene by alternative splicing. The ubiquitous enzymes consist of heterodimers with distinct large subunits associated with a common small subunit ; the association of tissue-specific large subunits with a small subunit has not yet been demonstrated. The large subunits of calpains can be subdivided into 4 protein domains. Domains I and III, whose functions remain unknown, show no homology with known proteins. Domain I, however, seems important for the regulation of the proteolytic activity. Domain II shows similarity with other cysteine proteases, sharing histidine, cysteine and asparagine residues at its active sites. Domain IV comprises four EF-hand structures which are potential calcium binding sites. In addition, three unique regions with no known homology are present in the muscle-specific nCL1 protein, namely NS, IS1 and IS2, the latter containing a nuclear translocation signal. These regions may be important for the muscle specific function of nCL1.

It is usually accepted that muscular dystrophies are associated with excess or deregulated calpains, and all the known approaches for curing these diseases are the use of antagonists of these proteases ; examples are disclosed in EP 359309 or EP 525420.

The invention results from the finding that, on the opposite to all these hypothesis, the LGMD2 disease is strongly correlated to the defect of a calpain which is expressed in healthy people.

The invention relates to the nucleic acid sequence such as represented in Figure 2 coding for a Ca<sup>++</sup> dependent protease, or calpain, which is involved in LGMD2 disease, and more precisely LGMD2A. It also relates to a part of this sequence provided it is able to code for a protein having a calcium-dependent protease activity involved in LGMD2, or a sequence derived from one of the above sequences by substitution, deletion or addition of one or more nucleotides provided that said sequence is still coding for said protein, all the nucleic acids yielding a sequence complementary to a sequence as defined above.

The genomic organisation of the human nCL1 gene has been determined by the inventors, and consists of 24 exons and extends over 40 kb as represented in Figure 8, and is also a part of the invention. About 35 kb of this gene have been sequenced. A systematic screening of this gene in LGMD2A families led to 5 the identification of 14 different mutations, establishing that a number of independent mutational events in nCL1 are responsible for LGMD2A. Furthermore, this is the first demonstration of a muscular dystrophy resulting from an enzymatic rather than a structural defect.

In the present specification, CANP3 means the protein which is a Ca<sup>++</sup> 10 dependent protease, or calpaïn, and coded by the nCL1 gene on chromosome 15.

The invention relates also to a protein, called CANP3, consisting in the amino acid sequence such as represented in figure 2 and which is involved, when mutated, in the LGMD2 disease.

15 The cDNA of the gene coding for CANP3, which is coding for the protein, is also represented in Figure 2, and is a part of the invention.

The protein coded by this DNA is CANP3, a calcium-dependent protease belonging to the Calpaïn family.

Are also included in the present invention the nucleic acid sequences 20 derived from the cDNA of Figure 2 by one or more substitutions, deletions, insertions, or by mutations in 5' or 3' non coding regions or in splice sites, provided that the translated protein has the protease, calcium-dependent activity, and when mutated, induce LGMD2 disease.

The nucleic acid sequence encoding the protein might be DNA or RNA and 25 be complementary to the nucleic acid sequence represented in Figure 2.

The invention also relates to a recombinant vector including a DNA sequence of the invention, under the control of a promoter allowing the expression of the calpaïn in an appropriate host cell.

A prokaryotic or eucaryotic host cell transformed by or transfected with a 30 DNA sequence comprising all or part of the sequence of Figure 2 is a part of the invention.

Such a host cell might be either :

- a cell which is able to secrete the protein and, this recombinant protein might be used as a drug to treat the LGMD2, or
  - a packaging cell line transfected by a viral or retroviral vector ; the cell lines bearing recombinant vector might be used as a drug for gene therapy of
- 5    LGMD2.

All the systems used today for gene therapy including adenoviruses and retroviruses and others described for example in « l'ADN médicament », (John Libbey, Eurotext, 1993), and bearing one of the DNA sequence of the invention are included herein by reference.

10      The examples hereunder and attached figures indicate how the structure of the gene was established, and how relationship between the gene and the LGMD was established.

Legend of the figures :

15      Figure 1:

A) Genomic organisation of the nCL1 gene

The gene covers a 40 kb region of which 35 were sequenced (Accession number pending). Introns and exons are drawn to scale, the latter being indicated by numbered vertical bars. The first intron is the largest one and  
20      remains to be fully sequenced. Position of intragenic microsatellites are indicated by asterisks. Arrows indicate the orientation of Alu (closed) and of Mer2 (greyed) repeat sequences.

B) *EcoRI* restriction map

An *EcoRI* (E) restriction map of this region was established with the help  
25      of cosmids from this region. The location of nCL1 gene is indicated as a black bar. The size of the corresponding fragments are indicated and are underlined when determined by sequence analysis.

C) Cosmid map of the nCL1 gene region.

Cosmids were from a cosmid library constructed by subcloning YAC  
30      774G4 (Richard in preparation) and are presented as lines. Dots on lines indicate positive STSs (indicated in boxed rectangles). A minimum of three cosmids cover the entire gene. T3, T7

Figure 2: Sequence of the human nCL1 cDNA (B) , and the flanking 5' (A) and 3' (C) genomic regions.

- A) and C) The polyadenylation signal and putative CAAT, TATAA sites are boxed. Putative Sp1 (position -477 to -472), MEF2 binding sites (-364 to -343) and CArG box (-685 to -672) are in bold. The Alu sequence present in the 5' region is underlined.
- B) The corresponding amino acids are shown below the sequence. The coding sequence between the ATG initiation codon and the TGA stop codon is 2466 bp, encoding for a 821 amino acid protein. The adenine in the first methionine codon has been assigned position 1. Locations of introns within the nCL1 gene are indicated by arrowheads. Nucleotides which differ from the previously published ones are indicated by asterisks.

Figure 3: Alignments of amino acid sequences of the muscle-specific calpains.

The human nCL1 protein is shown on the first line. The 3 muscle-specific sequences (NS, IS1 and IS2) are underlined. The second line corresponds to the rat sequence (Accession no P). The third and fourth lines show the deduced amino acid sequences encoded by pig and bovine Expressed Sequences Tagged (GenBank accession no U05678 and no U07858, respectively). The amino acids residues which are conserved among all known members of the calpains are in reverse letters. A period indicates that the same amino acid is present in the sequence. Letters refer to the variant amino acid found in the homologous sequence. Position of missense mutations are given as numbers above the mutated amino acid.

Figure 4: Distribution of the mutations along nCL1 protein structure.

- A) Positions of the 23 introns are indicated by vertical bars in relation to the corresponding amino acid coordinates.
- B) The nCL1 protein is depicted showing the four domains (I, II, III, IV) and the muscle specific sequences (NS, IS1 and IS2). The position of missense mutations within nCL1 domain are indicated by black dots. The effect of nonsense and frameshift mutations are illustrated as truncated lines, representing the extent of protein synthesised. Name of the corresponding families are indicated on the left of the line. The out of frame ORF is given by hatched lines.

Figure 5: Northern blot hybridisation of a nCL1 clone

A mRNA blot (Clontech) containing 2 µg of poly(A)+ RNA from each of eight human tissues was hybridised with a nCL1 genomic clone spanning exons 20 and 21. The latter detects a 3.6 kb mRNA present only in a line corresponding to the skeletal muscle mRNA.

Figure 6: Representative mutations identified by heteroduplex analysis.

Examples of mutation screening by heteroduplex analysis. Pedigree B505 shows the segregation of two different mutations in exon 22.

Figure 7: Homozygous mutations in the nCL1 gene

10 Detection by sequencing of mutations in exons 2 (a), 8 (b), 13 (c) and 22 (d). Sequences from a healthy control are shown above each mutant sequence. Asterisks indicate the position of the mutated nucleotides. The consequences on codon and amino acid residues are indicated on the left of the figure together with the name of the family.

15 Figure 8: Structure of nCL1 gene

Figure 8A represents the 5' part of the gene with exon 1.

Figure 8B represents the part of the gene including exons 2 to 8,

Figure 8C represents the part of the gene including exon 9,

Figure 8D represents the part of the gene including exons 10 to 24 including the 3' non transcribed region.

EXAMPLES

EXAMPLE 1

**Localisation of the nCL1 within the LGMD2A interval**

Detailed genetic and physical maps of the LGMD2A region were constructed (Fougerousse et al., 1994), following the primary linkage assignment to 15q (Beckmann et al., 1991). The disease locus was bracketed between the D15S129 and D15S143 markers, defining the cytogenetic boundaries of the LGMD2A region as 15q15.1-15q21.1 (Fougerousse et al., 1994). Construction and analysis of a 10-12 Mb YAC contig (Fougerousse et al., 1994) permitted us to map 33 polymorphic markers within this interval and to further narrow the LGMD2A region to between D15S514 and D15S222.

The nCL1 gene had been localised to chromosome 15 by hybridisation with sorted chromosomes and by Southern hybridisation to DNA from human-mouse cell hybrids (Ohno et al., 1989). cDNA capture using YACs from the LGMD2A interval allowed the identification of thirteen positional candidate genes. nCL1 was one of the two transcripts identified that showed muscle-specific expression as evidenced by northern blot analysis. The localisation was further confirmed by STS (for Sequence Tagged Site) assays. Primers used for the localisation of the nCL1 gene are P94in2, P94in13 and pcr6a3, as shown in Figure 1 and their characteristics being defined in Table 1.

10 Table 1: PCR primers used for localisation of the nCL1 gene.

Primer name	Primer sequence (5'-3')	Position within the cDNA	Annealing temp (°C)	PCR product size on	
				cDNA	genomic DNA
P94in2	ATGGAGCCAACAGAACTGA C GTATGACTCGGAAAAGAAC GT	341-360 428-448	58	108	1758
P94in13	TAAGCAAAAGCAGTCCCCA C TTGCTGTTCCCTCACTTCCCT G	1893-1912 1936-1956	58	64	1043
P94-6a3	GTTTCATCTGCTGCTTCGTT CTGGTTCAAGGCATACATGG T	2342-2361 2452-2471	56	130	818
P94ex1ter	TTCTTTATGTGGACCCTGAG TT ACGAACGGATGGGAACT	218-239 275-293	55	76	76

These primers are designed from different parts of the published human cDNA sequence (Sorimachi et al., 1989), and were used for an STS content screening on DNA from three chromosome 15 somatic cell hybrids and YACs from the LGMD2A contig. The results positioned the gene in a region previously defined as 15q15.1-q21.1 and on 3 YACs (774G4, 926G10, 923G7) localised in this region. The relative positions of STSs along the LGMD2A contig allowed to localise the gene between D15S512 and D15S488, in a candidate region suggested by linkage disequilibrium studies.

20 The same primers as above were used to screen a cosmid library from YAC 774G4. A group of 5 cosmids was identified (Fig. 1). Experiments with another nCL1 primer pair (P94ex1ter; Table 1) established that these cosmids cover all nCL1 exons except number 1, and that a second group of 4 cosmids contain this

exon (Fig. 1). A minimal set of three overlapping cosmids (2G8-2B11-1F11) covers the entire gene (Figure 1). DNA from these cosmids was used to construct an *EcoRI* restriction map of this region (Figure 1B).

#### EXAMPLE 2

5

##### **Determination of the nCL1 gene sequence**

Most of the sequences were obtained through shotgun sequencing of partial digests of cosmid 1F11 subcloned in M13 and bluescript vectors, and by walking with internal primers. The sequence assembly was made using the XBAP software of the Staden package (Staden) and was in agreement with the 10 restriction map of the cosmids. Sequences of exon 1 and adjacent regions were obtained by sequencing cosmid DNA or PCR products from human genomic DNA. The first intron is still not fully sequenced, but there is evidence that it may be between 10 to 16 kb in length (based on hybridisation of restriction fragments; data not shown). The entire gene, including its 5' and 3' regions, is more than 40 15 kb long, and shown in Figure 8.

###### a) the cDNA sequence

The used technology allows the implementation of the published human cDNA sequence of nCL1 (Sorimachi 1989). It contains the missing 129 bases corresponding to the N-terminal 43 amino acids (Figure 2). It also differs from it 20 at 12 positions. Three of which occur at third base positions of codons and preserve the encoded amino acid sequence. The other 9 differences lead to changes in amino-acid composition (Figure 2). As these different exons were sequenced repeatedly on at least 10 distinct genomes, we are confident that the sequence of Fig. 2 represents an authentic sequence and does not contain 25 minor polymorphic variants. Furthermore, these modifications increase the local similarity with the rat nCL1 amino acid sequence (Sorimachi), although the overall similarity is still 94 %.

The ATG numbered 1 in Figure 2 is the translation initiation site based on homology with the rat nCL1, and is within a sequence with only 5 nucleotides out 30 of 8 in common with the Kosak consensus sequence (Kosak M, 1984). Putative CCAAT and TATA boxes were observed 590, 324, (CCAAT) and 544 or 33 bp (TATA) upstream of the initiating ATG codon, respectively (Bucher, 1990). A GC-box binding the Sp1 protein (Dynan et al., 1983) was identified at position -477.

Consensus sequences corresponding to potential muscle-specific regulatory elements were identified (Fig. 2). These include a myocyte-specific enhancer-binding factor 2 (MEF2) binding site (Cserjesi P. 1991), a CArG box (Minty A. 1986) and 6 E-boxes (binding sites for basic Helix-Loop-Helix proteins frequently found in members of MyoD family; Blackwell et Weintraub, 1990). The functional significance of these putative transcription factor binding sites in the regulation of nCL1 gene expression remains to be established.

Two potential AAUAAA polyadenylation signals, were identified 520 and 777 bp downstream of the TGA stop codon. The sequencing of a partial nCL1 cDNA containing a polyA tail, demonstrated that the first AAUAAA is the polyadenylation signal. The latter is embedded in a region well conserved with the rat nCL1 sequence and is followed after 4 bp by a G/T cluster, present in most genes 3' of the polyadenylation site (Birnstiel et al., 1985). The 3'-untranslated region of the nCL1 mRNA is 565 bp long. The predicted length of the cDNA should therefore be approximately 3550 or 3000 bp.

b) Comparison with calpain

The sequence of the human nCL1 gene was compared to those of other calpains thereof (Figure 3). The most telling comparisons are with the homologous rat (Accession no J05121), bovine (Accession no U07858) and porcine (Accession no U05678) sequences. The accession numbers refers to those of international genebanks, such as GeneBank (N.I.H.) or EMBL Database (EMBL, Heidelberg). High local similarities between the human and rat DNA sequences are even observed in the 5' (75%) or in different parts of the 3' untranslated regions (over 60%) (data not shown). The high extent of sequence homology manifested by the human and rat nCL1 gene in their untranslated regions is suggestive of evolutionary pressures on common putative regulatory sequences.

c) Genomic organisation of the nCL1 gene

A comparison of the published nCL1 human cDNA (Sorimachi et al., 1989) with the corresponding genomic sequence led to the identification of 24 exons ranging in length from 12 bp (exon 13) to 309 bp (exon 1), with a mean size of 100 bp (Figure 1). The size of introns ranges from 86 bp to about 10-16 kb for intron 1.

The intron-exon boundaries as shown in Table 2 exhibit close adherence to 5' and 3' splice site consensus sequences (Shapiro and Senapathy, 1987).

Table 2: Sequences at the intron-exon junctions. A score expressing adherence to the consensus was calculated for each site according to Shapiro and Senapathy (1987). Sequences of exons and introns are in upper and lower cases, respectively. Size of exons are given in parenthesis.

splice donor site	score (%)	Intron	score (%)	splice acceptor site	Exon
...CTCCGgtgagt...	88.5	<-Intron 1->	99.0	...tttttgttacagGAAAT...	Exon 1 (309 bp) ->
...GCTAGgttagga...	83.5	<-Intron 2->	90.0	...gtgtctgcctgcagGGGAC...	Exon 2 (70 bp) ->
...TCCAGgtgagg...	92	<-Intron 3->	81.5	...acgcttcgtgcagTTCTG...	Exon 3 (119 bp) ->
...GCTAAgttaagc...	82	<-Intron 4->	81.5	...atccctctctaagGCTCC...	Exon 4 (134 bp) ->
...TTGATgttaagt...	87	<-Intron 5->	79.5	...ccatcgggcctcagGATGG...	Exon 5 (169 bp) ->
...CCCGGgtgtgt...	77.5	<-Intron 6->	91	...ttactgctctacagACAAT...	Exon 6 (144 bp) ->
...ATGAGgttaagc...	94	<-Intron 7->	78.5	...tctgtgtcttaagGTCCC...	Exon 7 (84 bp) ->
...GATAGgttaggt...	89	<-Intron 8->	91.5	...cattttcccaccagATGGA...	Exon 8 (86 bp) ->
...TTCTGgtgagt...	88	<-Intron 9->	92	...ttccaacctctcagGATGT...	Exon 9 (78 bp) ->
...CCCAGgtggga...	80	<-Intron 10->	68.5	...ttctgggggtcagATACT...	Exon 10 (161 bp) ->
...ACGAGgtgtgt...	85.5	<-Intron 11->	86	...tgtttcttcagGTTCC...	Exon 11 (170 bp) ->
...AAGAGgtatag...	70	<-Intron 12->	87	...tccccatcttcagATGCA...	Exon 12 (12 bp) ->
...TCTGAGtgagt...	76.5	<-Intron 13->	97	...tgtattcctcacagGGAAG...	Exon 13 (209 bp) ->
...CAGTGgtgagt...	89	<-Intron 14->	93.5	...cttttcttatgcagAAAAA...	Exon 14 (37 bp) ->
...CCAAGgttaggt...	89	<-Intron 15->	87	...cctccctctccagCCCAT...	Exon 15 (18 bp) ->
...CACAGgtgtct...	80	<-Intron 16->	88	...ttgtgcctccacagCCACA...	Exon 16 (114 bp) ->
...GAGATgtgagt...	84	<-Intron 17->	92.5	...cccttccctcctcagGACAT...	Exon 17 (78 bp) ->
...CAAACgtgagt...	83	<-Intron 18->	90	...ctccatccccccagACAAG...	Exon 18 (58 bp) ->
...TGGATgtatcc...	56	<-Intron 19->	88	...cctccctccctccagACAGA...	Exon 19 (65 bp) ->
...GGCAGgtggga...	80	<-Intron 20->	94	...tttcttatgcagAAATA...	Exon 20 (69 bp) ->
...CGCAGgtgtcg...	66	<-Intron 21->	91	...ggccccctccacagGATT...	Exon 21 (79 bp) ->
					Exon 22 (117 bp) ->

...GTTCAGtaagt...	79	<-Intron 22->	93.5	...gcattcttcacagGAGCT...	Exon 23 (59 bp) ->
<u>...TGGAGgtaaag...</u>	<u>81</u>	<u>&lt;-Intron 23-&gt;</u>	<u>79</u>	<u>...gggacttcttcagTGCT...</u>	<u>Exon 24 (27 bp) -&gt;</u>

When the genomic sequence was submitted to GRAIL analysis (Uberbacher et al., 1991), 11 exons were correctly recognised, 4 were not identified, 6 were inadequately defined and 2 were too small to be recognised (data not shown).

As already noted, the nCL1 gene has three unique sequence blocks, NS (amino acid residues 1 to 61), IS1 (residues 267 to 329) and IS2 (residues 578 to 653). It is interesting to note that each of these sequences, as well as the nuclear translocation signal inside IS2, are essentially flanked by introns (Fig. 4). The exon-intron organisation of the human nCL1 is similar to that reported for the chicken CANP (the only other large subunit calpain gene whose genomic structure is known; (Emori et al., 1986)).

Four microsatellite sequences were identified. Two of them are in the distal part of the first intron: an (AT)<sub>14</sub> and an previously identified mixed-pattern microsatellite, S774G4B8, which was demonstrated to be non polymorphic (Fougerousse et al., 1994). A (TA)<sub>7</sub>(CA)<sub>4</sub>(GA)<sub>13</sub> was identified in the second intron and genotyping of 64 CEPH unrelated individuals revealed two alleles (with frequencies of 0.10 and 0.90). The fourth microsatellite is a mixed (CA)<sub>n</sub>(TA)<sub>m</sub> repeat present in the 9th intron. The latter and the (AT)<sub>14</sub> repeat have not been investigated for polymorphism. Fourteen repetitive sequences of the Alu family and one Mer2 repeat were identified in the nCL1 gene (Fig. 1C), which has, thus, on the average one Alu element per 2.5 kb.

Southern blot experiments (Ohno et al., 1989) and STS screening (data not shown) suggest that there is but one copy per genome of this member of the calpain family.

### 25      EXAMPLE 3

#### Expression of the nCL1 gene

The pattern of tissue-specificity was investigated by northern blot hybridisation with a genomic subclone probe from cosmid 1F11 spanning exons 20 and 21. There is no evidence for the existence of an alternatively spliced form 30 of nCL1, although this cannot be excluded. A transcript of about 3.4-3.6 kb was

detected in skeletal muscle mRNA (Figure 5). This size therefore favours that the position -544 is the functional TATA box.

Transcription studies suggested that it is an active gene rather than a pseudogene and its muscle-specific pattern of expression is consistent with the 5 phenotype of this disorder (Sorimachi et al., 1989 and Figure 5).

#### EXAMPLE 4

##### **Mutation screening**

nCL1 fulfils both positional and functional criteria to be a candidate gene for LGMD2A. To evaluate its role in the etiology of this disorder, nCL1 was 10 systematically screened in 38 LGMD2 families for the presence of nucleotide changes using a combination of heteroduplex (Keen et al., 1991) and direct sequence analyses.

PCR primers were designed to specifically amplify the exons and splice junctions and also the regions containing the putative CAT, TATA boxes and the 15 polyadenylation signal of the gene as shown in Table 3.

**Table 3:** PCR primers used for the analysis of the nCL1 gene in LGMD patients.

amplified region	Primer sequences (5'-3')	Size (bp)	Annealing temp. (°C)
promotor	TTCAGTACCTCCGTTCA GATGCTTGAGCCAGGAAAC	296	59
exon 1	CTTCCTTGAAGGTAGCTGTAT GAGGTGCTGAGTGAGAGGAC	438	60
exon 2	ACTCCGTCTCAAAAAAACCT ATTGTCCCTTTACCTCCTGG	239	57
exon 3	TGGAAGTAGGAGAGTGGGCA GGGTAGATGGGTGGGAAGTT	354	58
exon 4	GAGGAATGTGGAGGAAGGAC TTCCTGTGAGTGAGGTCTCG	292	59
exon 5	GGAACTCTGTGACCCCAAAT TCCTCAAACAAAATTCGC	325	56
exon 6	GTTCCCTACATTCTCCATCG GTTATTCAACCCAGACCCCTT	315	57
exon 7	AATGGGTTCTCTGGTTACTGC AGCACGAAAGCAAAGATAAA	333	56
exon 8	GTAAGAGATTGCCCCCAG TCTGCGGATCATTGGTTTG	321	58
exon 9	CCTTCCCTTCTCCTGCTTC CTCTCTTCCCCACCCCTTACC	173	56
exon 10	CCTCCTCACCTGCTCCCATA TTTTTCGGCTTAGACCCCTCC	251	56
exon 11	TGTGGGAATAGAAATAATGG CCAGGAGCTCTGTGGGTCA	355	57
exon 12	GGCTCCTCATCCTCATTACA GTGGAGGAGGGTGAGTGTGC	312	61
exon 13	TGTGGCAGGACAGGACGTT	337	60

		TTCAACCTCTGGAGTGGGCC		
exon 14		CACCAGAGCAAACCGTCCAC	230	61
		ACAGCCCAGACTCCCATTCC		
exon 15		TTCTCTTCTCCCTTCACCC	225	57
		ACACACTTCATGCTCTACCC		
exon 16		CCGCCTATTCCCTTCCTCTT	331	56
		GACAAACTCCTGGGAAGCCT		
exon 17		ACCTCTGACCCCTGTGAACC	270	61
		TGTGGATTGTGTGCTACGC		
exon 18		CATAAATAGCACCGACAGGG	258	59
		GGGATGGAGAACAGTGAGGA		
exon 19		TCCCTCACTTTCTCCATCCC	159	57
		ACCCCTGTATGTTGCCTTGG		
exons 20-21		GGGGATTTGCTGTGTGCTG	333	61
		ATTCCTGCTCCCACCGTCTC		
exon 22		CACAGAGTGTCCGAGAGGCA	282	57
		GGAGATTATCAGGTGAGATGCC		
exons 22-23		CAGAGTGTCCGAGAGGCAGGG	608	61
		CGTTGACCCCTCCACCTTGA		
exon 24		GGGAAAACATGCACCTTCTT	375	58
		TAGGGGGTAAAATGGAGGAG		
polyadenylation signal		ACTAACTCAGTGGAAATAGGG	413	56
		GGAGCTAGGATAGCTCAAT		

PCR products made on DNA from blood of specific LGMD2A patients were then subjected either to heteroduplex analysis or to direct sequencing, depending on whether the mutation, based on haplotype analysis, was expected to be homozygous or heterozygous, respectively. It was occasionally necessary to clone the PCR products to precisely identify the mutations (i.e., for microdeletions or insertions and for some heterozygotes). Disease-associated mutations are summarised in Table 4 hereunder and their position along the protein is shown in Fig. 4.

**Table 4:** nCL1 mutations in LGMD2A families.

Codons and amino acid positions are numbered on the basis of the cDNA sequence starting from ATG.

Exon	Families	Nucleotide position	Nucleotide change	Amino acid position	Amino acid change	Restriction si
2	B519*	328	<u>C</u> GA-> <u>T</u> GA	110	Arg->stop	
4	M42	545	<u>C</u> TG -> <u>C</u> AG	182	Leu->Gln	
4	M1394; M2888	550	CAA -> CA	184	frameshift	
5	M35; M37	701	<u>GG</u> G -> <u>G</u> AG	234	Gly->Glu	

6	M32	945	CGG -> CG	315	frameshift	-SmaI
8	M2407*	1061	G <u>T</u> G -> G <u>GG</u>	354	Val-> Gly	
8	M1394	1079	T <u>GG</u> -> T <u>AG</u>	360	Trp->stop	-BstnI, -Eco
11	M2888	1468	C <u>GG</u> -> T <u>GG</u>	490	Arg->Trp	
13	R12*	1715	C <u>GG</u> -> C <u>A</u> G	572	Arg->Gln	-MspI
19	R27	2069-2070	deletion AC	690	frameshift	
21	R14; R17	2230	A <u>GC</u> -> G <u>GC</u>	744	Ser->Gly	-AluI
22	A*; B501*; M32	2306	C <u>GG</u> -> C <u>A</u> G	769	Arg->Gln	
22	B505	2313-2316	deletion AGAC	771-772	frameshift	
22	R14; B505	2362-2363	AG -> TCATCT	788	frameshift	

The first letter of the family code refers to the origin of the population B= Brazil, M= metropolitan France, R = Isle of La Réunion, A= Amish.

Each mutation was confirmed by heteroduplex analysis, by sequencing of both strands in several members of the family or by enzymatic digestion when the mutation resulted in the modification of a restriction site. Segregation analyses of the mutations, performed on DNAs from all available members of the families, confirmed that these sequence variations are on the parental chromosome carrying the LGMD2A mutation. To exclude the possibility that the missense substitutions might be polymorphisms, their presence was systematically tested in a control population: none of these mutations was seen among 120 control chromosomes from the CEPH reference families.

#### EXAMPLE 5:

#### **Analysis of families genes, chromosome-15 ascertained families**

The initial screening for causative mutations was performed on families, each containing a LGMD gene located on chromosome 15. These included families from the Island of La Réunion (Beckmann et al., 1991), from the Old Order Amish from northern Indiana (Young et al., 1992,) and 2 Brazilian families (Passos Bueno et al., 1993).

20      a) Reunion Island families

Genealogical studies and geographic isolation of the families from the Isle of La Réunion were suggestive of a single founder effect. Genetic analyses are,

however, inconsistent with this hypothesis as the families present haplotype heterogeneity. At least, six different carrier chromosomes are encountered, (with affected individuals in several families being compound heterozygotes). Distinct mutations corresponding to four of these six haplotypes have been identified thus far.

In family R14, exons 13, 21 and 22 showed evidence for sequence variation upon heteroduplex analysis (Fig. 6). Sequencing of the associated PCR products revealed (i) a polymorphism in exon 13, (ii) a missense mutation (A->G) in exon 21 transforming the Ser<sup>744</sup> residue to a glycine in the loop of the second EF-hand in domain IV of the protein (Figure 4), and (iii) a frameshift mutation in exon 22. The exon 21 mutation and the polymorphism in exon 13 form an haplotype which is also encountered in family R17. Subcloning of the PCR products was necessary to identify the exon 22 mutation. Sequencing of several clones revealed a replacement of AG by TCATCT (data not shown). This frameshift mutation causes premature termination at nucleotide 2400 where an in frame stop codon occurs (Figure. 4).

The affected individuals in family R12 are homozygous for all markers of the LGMD2A interval (Allamand, submitted). Sequencing of the PCR products of exon 13 revealed a G to A transition at base 1715 of the cDNA resulting in a substitution of glutamine for Arg<sup>572</sup> (Figure. 7) within domain III, a residue which is highly conserved throughout all known calpains. This mutation, detectable by loss of Mspl restriction site, is present only in this family and in no other examined LGMD2A families or unrelated controls.

In family R27, heteroduplex analysis followed by sequencing of the PCR products of an affected child revealed a two base pair deletion in exon 19 (Figure. 6 and table 4). One AC out of three is missing at this position of the sequence, producing a stop codon at position 2069 of the cDNA sequence (Figure 4).

b) Amish families

As expected, due to multiple consanguineous links, the examined LGMD2A Northern Indiana Amish patients were homozygous for the haplotype on the chromosome bearing the mutant allele (Allamand, submitted). A (G->A) missense mutation was identified at nucleotide 2306 within exon 22 (Fig. 7). The

resulting codon change is CGG to CAG, transforming Arg<sup>769</sup> to glutamine. This residue, which is conserved throughout all members of the calpain family in all species, is located in domain IV of the protein within the 3rd EF-hand at the helix-loop junction (ref). This mutation was encountered in a homozygous state  
5 in all patients from 12 chromosome 15-linked Amish families, in agreement with the haplotype analysis. We also screened six Southern Indiana Amish LGMD families, for which the chromosome 15 locus was excluded by linkage analyses (Allamand ESHG, submitted, ASHG 94). As expected, this nucleotide change was not present in any of the patients from these families, thus confirming the  
10 genetic heterogeneity of this disease in this genetically related isolate.

c) Brazilian families

As a result of consanguineous marriages, two Brazilian families (B501, B519) are homozygous for extended LGMD2A carrier haplotypes (data not shown). Sequencing PCR products from affected individuals of these families  
15 demonstrated that family B501 has the same exon 22 mutation found in northern Indiana Amish patients (Figure 7), but embedded in a completely different haplotype. In family B519, the patients carry a C to T transition in exon 2, replacing Arg<sup>328</sup> with a TGA stop codon (Figure 7), thus leading, presumably, to a very truncated protein (Figure 4).

20 d) Analysis of other LGMD families

Having validated the role of the candidate gene in the chromosome 15 ascertained families, we next examined by heteroduplex analysis LGMD families for which linkage data were not informative. These included one Brazilian (B505) and 13 metropolitan French pedigrees.

25 Heteroduplex bands were revealed for exons 1, 3, 4, 5, 6, 8, 11, 22 of one or more patients (Figure 6). Of all sequence variants, 10 were identified as possible pathogenic mutations (5 missense, 1 nonsense and 4 frameshift mutations) and 3 as polymorphisms with no change of amino acid of the protein. All causative mutations identified are listed in Table 4 here-above. Identical  
30 mutations were uncovered in apparently unrelated families. The mutations shared by families M35 and M37, and M2888 and M1394, respectively, are likely to be the consequence of independent events since they are embedded in different marker haplotypes. In contrast, it is likely that the point mutation in exon

22 of the Amish and in the M32 kindreds corresponds to the same mutational event as both chromosomes share a common four marker haplotype (774G4A1-774G4A10-774G454D-774G4A2) around nCL1 (data not shown), possibly reflecting a common ancestor. The same holds true for the AG to TCATCT substitution mutation encountered in exon 22 in families B505 and R14. The 5 exon 8 (T->G) transversion is present in the two carrier chromosomes of M2407, the only metropolitan family homozygous by haplotype, possibly reflecting an undocumented consanguinity. For some families, no disease-causing mutation has been detected thus far (M40 for example).

10 In addition to the polymorphism present in exon 13 in families R14 and R17 (position 668) and in the intragenic microsatellites, four additional neutral variations were detected: a (T->C) transition at position 96, abolishing a *Ddel* restriction site in exon 1 in M31; a (C->T) transition in exon 3 (position 495) in M40 and in M37 forming a haplotype with the exon 5 mutation (in the former 15 family, this polymorphism does not cosegregate with the disease); a (T->C) transition in the paternally derived promotor in M42 at position -428, which was also evidenced in healthy controls; and a variable poly(G) in intron 22 close to the splice site in families R20, R11, R19, M35 and M37. The latter is also present in the members of the CEPH families, but is not useful as a genetic 20 marker as the visualisation and interpretation of mononucleotide repeat alleles is difficult.

In total, sixteen independent mutational events representing fourteen different mutations were identified. All mutations cosegregate with the disease in LGMD2A families. The characterised morbid calpain alleles contain nucleotide 25 changes which were not found in alleles from normal individual. The discovery of two nonsense and five frameshift mutations in nCL1 supports the hypothesis that a deficiency of this product causes LGMD2A. All seven mutations result in a premature in-frame stop codon, leading to the production of truncated and presumably inactive proteins (Figure 4). Evidences for the morbidity of the 30 missense mutations come from (1) the relative high incidence of such mutations among LGMD2A patients ; although it is difficult in the absence of functional assays to differentiate between a polymorphism and a morbid mutation, the occurrence of different "missense" mutations in this gene cannot all be

accounted for as rare private polymorphisms; (2) the failure to observe these mutations in control chromosomes; and (3) the occurrence of mutations in evolutionarily conserved residues and/or in regions of documented functional importance. Four of seven missense mutations change an amino acid which is  
5 conserved in all known members of the calpain family in all species (Figure 3). Two of the remaining mutations affect less conserved amino acid residues, but are located in important functional domains. The substitution V354G in exon 8 is  
4 residues before the asparagine at the active site and S744G in exon 21 is within the loop of the second EF-hand and may impair the calcium-dependent  
10 regulation of calpain activity or the interaction with a small subunit (Figure 4). Several missense mutations change a hydrophobic residue to a polar one, or vice versa (Table 4) possibly disrupting higher order structures.

## METHODS

### Description of the patients

15 The LGMD2A families analysed were from 4 different geographic origins. They included 3 Brazilian families, 13 interrelated nuclear families from the Isle of la Réunion, 10 French metropolitan families and 12 US Amish families. The majority of these families were previously ascertained to belong to the chromosome 15 group by linkage analysis (Beckmann, 1991; Young, Passos-  
20 Bueno et al., 1993). However, some families from metropolitan France as well as one Brazilian family, B505, had non significant lodscores for chromosome 15. Genomic DNA was obtained from peripheral blood lymphocytes.

### Sequencing of cosmid c774G4-1F11 and EcoRI restriction map of cosmids.

25 Cosmid 1F11 (Figure 1C) was subcloned following DNA preparation through Qiagen procedure (Qiagen Inc., USA) and partial digestion with either Sau3A, RsaI or AluI. Size-selected restriction fragments were recovered from low-melting agarose and eventually ligated with M13 or Bluescript (Stratagene, USA) vectors. After electroporation in *E.coli*, recombinant colonies were picked in 100 µl of LB/ampicillin media. PCR reactions were performed on 1 µl of the culture in  
30 10 mM Tris-HCl, pH 9.0, 50 mM KCl, 1.5 mM MgCl<sub>2</sub>, 0.1% Triton X-100, 0.01 gelatine, 200µM of each dNTP, 1 U of Taq Polymerase (Amersham) with 100 ng of each vector primers. Amplification was initiated by 5 min denaturation at 95°C, followed by 30 cycles of 40 sec denaturation at 92°C and 30 sec annealing

at 50°C. PCR products were purified through Microcon devices (Amicon, USA) and sequenced using the dideoxy chain termination method on an ABI sequencer (Applied Biosystems, Foster City, USA). The sequences were analysed and alignments performed using the XBAP software of the Staden package, version 93.9 (Staden, 1982). Gaps between sequence contigs were filled by walking with internal primers. EcoRI restriction map of cosmids was performed essentially as described in Sambrook et al. (1989).

#### Northern Blot analysis

The probes were labelled by random priming with dCTP-(a<sup>32</sup>P).  
10 Hybridisation was performed to human multiple tissue northern blots as recommended by the manufacturer (Clontech, USA).

#### Analysis of PCR products from LGMD2A families

One hundred ng of human DNA were used per PCR under the buffer and cycle conditions described in Fougerousse (1994) (annealing temperature shown  
15 in Table 3). Heteroduplex analysis (Keene et al., 1991) was performed by electrophoresis of ten µl of PCR products on a 1.5 mm-thick Hydrolink MDE gels (Bioprobe) at 500-600 volt for 12-15 h depending of the fragment length. Migration profile was visualised under UV after ethidium bromide staining.

For sequence analysis, the PCR products were subjected to dye-dideoxy  
20 sequencing, after purification through microcon devices (Amicon, USA). When necessary, depending on the nature of the mutations (e.g., frameshift mutation or for some heterozygotes), the PCR products were cloned using the TA cloning kit from Invitrogen (UK). One µl of product was ligated to 25 ng of vector at 12°C overnight. After electroporation into XL1-blue bacteria, several independent  
25 clones were analysed by PCR and sequenced as described above.

The invention results from the finding that the nCL1 gene when it is mutated is involved in the etiology of LGMD2A. It is exactly the contrary to what is stated in the litterature, e.g. that the disease is accompanied by the presence of a deregulated calpaïn. Identification of nCL1 as the defective gene in LGMD2A  
30 represents the first example of muscular dystrophy caused by mutation affecting a gene which is not a structural component of muscle tissue, in contrast with previously identified muscular dystrophies such as Duchenne and Becker (Bonilla et al., 1988), severe childhood autosomal recessive (Matsumara et al.,

1992), Fukuyama (Matsumara et al., 1993) and merosin-deficient congenital muscular dystrophies (Tomé et al., 1994).

The understanding of the LGMD2A phenotype needs to take into account the fact that there is no active nCL1 protein in several patients, a loss compatible with the recessive manifestation of this disease. Simple models in which this protease would be involved in the degradation or destabilisation of structural components of the cytoskeleton, extracellular matrix or dystrophin complex can therefore be ruled out. Furthermore, there are no signs of such alterations by immunocytogenetic studies on LGMD2 muscle biopsies (Matsumara et al., 1993; Tomé et al., 1994). Likewise, since LGMD2A myofibers are apparently not different from other dystrophic ones, it seems unlikely that this calpain plays a role in myoblast fusion, as proposed for ubiquitous calpains (Wang et al., 1989).

All the data disclosed in these examples confirm that the nCL1 gene is a major gene involved in the disease when mutated.

The fact that morbidity results from the loss of an enzymatic activity raises hopes for novel pharmaco-therapeutic prospects. The availability of transgenic models will be an invaluable tool for these investigations.

The invention is also relative to the use of a nucleic acid or a sequence of nucleic acid of the invention, or to the use of a protein coded by the nucleic acid for the manufacturing of a drug in the prevention or treatment of LGMD2.

The finding that a defective calpain underlies the pathogenesis of LGMD2A may prove useful for the identification of the other loci involved in the LGMDs. Other forms of LGMD may indeed be caused by mutations in genes whose products are the CANP substrates or in genes involved in the regulation of nCL1 expression. Techniques such as the two-hybrid selection system (Fields et al., 1989) could lend themselves to the isolation of the natural protein substrate(s) of this calpain, and thus potentially help to identify other LGMD loci.

The invention also relates to the use of all or a part of the peptidic sequence of the enzyme, or of the enzyme, product of nCL1 gene, for the screening of the ligands of this enzyme, which might be also involved in the etiology and the morbidity of LGMD2.

The ligands which might be involved are for example substrate(s), activators or inhibitors of the enzyme.

The nucleic acids of the invention might also be used in a screening method for the determination of the components which may act on the regulation of the gene expression.

A process of screening using either the enzyme or a host recombinant cell, 5 containing the nCL1 gene and expressing the enzyme, is also a part of the invention.

The pharmacological methods, and the use of nucleic acid and peptidic sequences of the invention are very potent applications.

The methods used for such screenings of ligands or regulatory elements are 10 those described for example for the screening of ligands using cloned receptors.

The identification of mutations in the nCL1 gene provides the means for direct prenatal or presymptomatic diagnosis and carrier detection in families in which both mutations have been identified. Gene-based accurate classification of LGMD2A families should prove useful for the differential diagnosis of this 15 disorder.

The invention relates to a method of detection of a predisposition to LGMD2 in a family or a human being, such method comprising the steps of :

- selecting one or more exons or flanking sequences which are sensitive in said family;
- 20 - selecting the primers specific for the or these exons or their flanking sequences, a specific example being the PCR primers of Table 3, or an hybrid thereof,
- amplifying the nucleic acid sequence, the substrate for this amplification being the DNA of the human being to be checked for the predisposition, and
- 25 - comparing the amplified sequence to the corresponding sequence derived from Figure 2 or Figure 8.

Table 2 indicates the sequences of the introns-exons junctions, and primers comprising in their structure these junctions are also included in the invention.

All other primers suitable for such RNA or DNA amplification may be used in 30 the method of the invention.

In the same way, any suitable amplification method : PCR (for Polymerase Chain Reaction ®) NASBA ® (for Nucleic acid Sequence Based Amplification), or others might be used.

The methods usually used in the detection of one site mutations, like ASO (Allele specific PCR), LCR, or ARMS (Amplification Refactory Mutation System) may be implemented with the specific primers of the invention.

5 The primers, such as described in Tables 1 and 3, or including junctions of Table 2, or more generally including the flanking sequences of one of the 24 exons are also a part of the invention.

The kit for the detection of a predisposition to LGMD2 by nucleic acid amplification is also in the scope of the invention, such a kit comprises at least PCR primers selected from the group of :

- 10 a) in those described in table 1  
b) in those described in table 3  
c) those including the introns-exons junctions of Table 2.  
d) derived from primers defined in a), b) or c).

15 The nucleic acid sequence of claim 1 to 3 might be inserted in a viral or a retroviral vector, said vector being able to transfect a packaging cell line.

The packaging transfected cell line, might be used as a drug for gene therapy of LGMD2.

The treatment of LGMD2 disease by gene therapy is implemented by a pharmaceutical composition containing a component selected from the group of :

- 20 a) a nucleic acid sequence according to claims 1 to 4,  
b) a cell line according to claim 24,  
c) an aminoacid sequence according to claims 5 to 9.

## REFERENCES

- 5 Arikawa, E., Hoffman, E. P., Kaido, M., Nonaka, I., Sugita, H. and Arahata, K. (1991). The frequency of patients with dystrophin abnormalities in a limb-girdle patient population. *Neurology* 41, 1491-1496.
- 10 Bashir, R., Strachan, T., Keers, S., Stephenson, A., Mahjneh, I., Marconi, G., Nashef, L. and Bushby, K. M. D. (1994). A gene for autosomal recessive limb-girdle muscular dystrophy maps to chromosome 2p. *Hum. Mol. Genet.* 3, 455-457.
- 15 Beckmann, J. S., Richard, I., Hillaire, D., Broux, O., Antignac, C., Bois, E., Cann, H., Cottingham, R. W., Jr., Feingold, N., Feingold, J., Kalil, J., Lathrop, G. M., Marcadet, A., Masset, M., Mignard, C., Passos-Bueno, M. R., Pellerain, N., Zatz, M., Dausset, J., Fardeau, M. and Cohen, D. (1991). A gene for limb-girdle muscular dystrophy maps to chromosome 15 by linkage. *C. R. Acad. Sci. Paris. III* 312, 141-148.
- 20 Birnstiel, M. L., Busslinger, M. and Sturb, K. (1985). Transcription termination and 3' processing: The end is in site! *Cell* 41, 349-359.
- 25 Blackwell, T. K. and Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* 250, 1104-1110.
- Bonilla, E., Samitt, C. E., Miranda, A. F., Hays, A. P., Salviati, G., DiMauro, S., Kunkel, L. M., Hoffman, E. P. and Rowland, L. P. (1988). Duchenne muscular dystrophy: deficiency of dystrophin at the muscle cell surface. *Cell* 54, 447-452.
- 30 Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* 212, 563-578.

- Bushby, K. M. D. (1994). Limb-girdle muscular dystrophy. In Diagnostic criteria for neuromuscular disorders, A. E. H. Emery, ed. (Baarn, The Netherlands: ENMC), pp 25-31.
- 5 Croall, D. E. and Demartino, G. N. (1991). Calcium-activated neutral protease (calpain) system: structure, function, and regulation. *Physiol. Rev.* 71, 813-847.
- Dynan, W. S. and Tjian, R. (1983). The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* 35, 79-87.
- 10 Emery, A. E. H. (1991). Population frequencies of inherited neuromuscular diseases - a world survey. *Neuromuscular Disorders* 1, 19-29.
- Emori, Y., Ohno, S., Tobita, M. and Suzuki, K. (1986). Gene structure of calcium-dependent protease retains the ancestral organization of the calcium-binding protein gene. *FEBS lett.* 194, 249-252.
- 15 Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-246.
- Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* 340, 245-246.
- 20 Fougerousse, F., Broux, O., Richard, I., Allamand V., Pereira de Souza, A., Bourg N., Brenguier L., Devaud C., Pasturaud P., Roudaut C., Chiannilkulchai N., Hillaire D., Bui H., Chumakov I., Weissenbach J., Cherif D., Cohen D. and J. S. Beckmann (1994). Mapping of a chromosome 15 region involved in Limb-Girdle Muscular Dystrophy. *Hum. Mol. Genet.* 3, 285-293.
- 25 Goll, D. E., Thompson, V. F., Taylor, R. G. and Zalewska, T. (1992). Is Calpain activity regulated by membranes and autolysis or by calcium and calpastatin? *BioEssays* 14, 549-556.
- Gosset, L. A., Kelvin, D. J., Sternberg, E. A. and Olson, E. (1989). A new myocyte-specific enhancer-binding factor that recognizes a conserved element associated with multiple muscle-specific genes. *Mol. Cell. Biol.* 9, 5022-5033.

- Hirai, S., Kawasaki, H., Yaniv, M. and Suzuki, K. (1991). Degradation of transcription factors, c-Jun and c-Fos, by calpain. FEBS lett. 1, 57-61.
- 5 Imajoh, S., Kawasaki, H. and Suzuki, K. (1986). Limited autolysis of calcium-activated neutral protease (CANP): reduction of the Ca<sup>2+</sup> requirement is due to the NH<sub>2</sub>-terminal processing of the large subunit. J. Biochem. 100, 633-642.
- Jackson, C. E. and Carey, J. H. (1961). Progressive muscular dystrophy: 10 autosomal recessive type. Pediatrics 77-84.
- Keen, J., Lester D., Inglehearn, C., Curtis, A. and Bhattacharya, S. (1991). Rapid detection of single base mismatches as heteroduplexes on Hydrolink gels. Trends Genet. 7, 5.
- 15 Kosak, M. (1984). Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. Nucleic Acids Res. 12, 857-872.
- Lovett, M., Kere, J. and Hinton, L. M. (1991). Direct selection: a method for the 20 isolation of cDNAs encoded by large genomic regions. Proc. Natl. Acad. Sci. USA 88, 9628-9632.
- Matsumara, K., Tomé F. M. S., Collin H., Azibi K., Chaouch M., Kaplan J-K., Fardeau M. and Campbell K., P. (1992). Deficiency of the 50K dystrophin- 25 associated glycoprotein in severe childhood autosomal recessive muscular dystrophy. Nature 359, 320-322.
- Matsumura, K., Nonaka, I. and Campbell, K. P. (1993). Abnormal expression of dystrophin-associated proteins in Fukuyama-type congenital muscular dystrophy. 30 Lancet 341, 521-522.

- Minty, A. and Kedes, L. (1986). Upstream regions of the human cardiac actin gene that modulate its transcription in muscle cells: presence of an evolutionarily conserved repeated motif. *Mol. Cell. Biol.* 6, 2125-2136.
- 5 Miyamoto, S., Maki, M., Schmitt, M. J., Hatanaka, M. and Verma, I. M. (1994). TNF-a- induced phosphorylation of IKB is a signal for its degradation but not dissociation from NF-KB. *Proc. Natl. Acad. Sci. USA* *in press*.
- Morton, N. E. and Chung, C. S. (1959). Formal genetics of muscular dystrophy.
- 10 Am. J. Hum. Genet. 11, 360-379.
- Murachi, T. (1989). Intracellular regulatory system involving calpain and calpastatin. *Biochemistry Int.* 18, 263-294.
- 15 Ohno, S., Emori, Y., Imajoh, S., Kawasaki, H., Kisaragi, M. and Suzuki, K. (1984). Evolutionary origin of a calcium-dependent protease by fusion of genes for a thiol protease and a calcium-binding protein? *Nature* 312, 566-570.
- Ohno, S., Minoshima, S., Kudoh, J., Fukuyama, R., Shimizu, Y., Ohmi-Imajoh, S.,
- 20 Shimizu, N., Suzuki, K. (1989). Four genes for the calpain family locate on four different chromosomes. *Cytogen. Cell Genet.* 51, 1054.
- Passos-Bueno, M.-R., Richard, I., Vainzof, M., Fougerousse, F., Weissenbach, J., Broux, O., Cohen, D., Akiyama, J., Marie, S. K. N., Carvalho, A. A.,
- 25 Guilherme, L., Kalil, J., Tsanaclis, A. M., Zatz, M. and Beckmann, J. S. (1993). Evidence of genetic heterogeneity in the autosomal recessive adult forms of limb-girdle muscular dystrophy following linkage analysis with 15q probes in Brazilian families. *J. Med. Genet.* 30, 385-387.
- Richard, I., Broux, O., Chiannilkulchai, N., Fougerousse, F., Allamand, V., Bourg, N., Brenguier, L., Devaud, C., Pasturaud, P., Roudaut, C., Lorenzo, F., Sebastiani-Kabatchis, C., Schultz, R. A., Polymeropoulos, M. H., Gyapay, G.,

- Auffray, C. and Beckmann, J. (1994). Regional localization of human chromosome 15 loci. *Genomics* 23, 619-627.
- Sambrook, J., Fritsh, E. F. and Maniatis, T. (1989). Molecular cloning: a laboratory manual. Cold spring Harbor Laboratory Press, Cold spring Harbor, USA.
- Shapiro, M. and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* 15, 7155-7174.
- Sorimachi, H., Imajoh-Ohmi, S., Emori, Y., Kawasaki, H., Ohno, S., Minami, Y. and Suzuki K. (1989). Molecular cloning of a novel mammalian calcium-dependant protease distinct from both m- and mu- type. Specific expression of the mRNA in skeletal muscle. *J. Biol. Chem.* 264, 20106-20111.
- Sorimachi, H., Ishiura, S. and Suzuki, K. (1993a). A novel tissue-specific calpain species expressed predominantly in the stomach comprises two alternative splicing products with and without  $\text{Ca}^{2+}$ -binding domain. *J. Biol. Chem.* 268, 19476-19482.
- Sorimachi, H., Toyama-Sorimachi, N., Saido, T. C., Kawasaki, H., Sugita, H., Miyasaka, M., Arahata, K., Ishiura, S. and Suzuki, K. (1993b). Muscle-specific calpain, p94, is degraded by autolysis immediately after translation, resulting in disappearance from muscle. *J. Biol. Chem.* 268, 10593-10605.
- Staden, R. (1982). An interactive graphic program for comparing and aligning nucleic acid and amino acid sequences. *Nucleic Acids Res.* 10, 2951-2961.
- Suzuki, K. and Ohno, S. (1990). Calcium activated neutral protease. Structure-function relationship and functional implications. *Cell Struct. Funct.* 15, 1-6.

- Tagle, D. A., Swaroop, M., Lovett, M. and Collins, F. S. (1993). Magnetic bead capture of expressed sequences encoded within large genomic segments. *Nature* 361, 751-753.
- 5 Tomé, F. M. S., Evangelista T., Leclerc A., Sunada Y., Manole E., Estournet B., Barois A., Campbell K. P. and Fardeau M. (1994). Congenital muscular dystrophy with merosin deficiency. *C. R. Acad. Sci. Paris* 317, 351-357.
- Uberbacher, E. C. and Mural, R. J. (1991). Locating protein-coding regions in  
10 human DNA sequences by a multiple sensor-neural network approach. *Proc. Natl. Acad. Sci. USA* 88, 11261-11265.
- Walton, J. N. and Nattrass, F. J. (1954). On the classification, natural history and treatment of the myopathies. *Brain* 77, 169-231.  
15
- Wang, K. W., Villalobo, A. and Roufogalis, B. D. (1989). Calmodulin-binding proteins as calpain substrates. *Biochem. J.* 262, 693-706.
- Young, K., Foroud, T., Williams, P., Jackson, C. E., Beckmann, J. S., Cohen, D.,  
20 Conneally, P. M., Tischfield, J. and Hodes, M. E. (1992). Confirmation of linkage of limb-girdle muscular dystrophy, type-2, to chromosome 15. *Genomics* 13, 1370-1371.

CLAIMS

- 5
1. A nucleic acid sequence comprising :
    - 1) the sequence represented in Figure 8; or
    - 2) the sequence represented in Figure 2; or
    - 5        3) a part of the sequence of Figure 2 with the proviso that it is able to code for a protein having a calcium dependant protease activity involved in a LGMD2 disease ; or
    - 10      4) a sequence derived from a sequence defined in 1), 2) or 3) by substitution, deletion or addition of one or more nucleotides with the proviso that said sequence still codes for said protease.
  2. A nucleic acid sequence that is complementary to a nucleic acid sequence according to claim 1.
  3. A nucleic acid sequence comprising in its structure a nucleotidic sequence according to claim 1 or 2, under the control of regulatory elements, 15 and involved in the expression of calpain activity in a LGMD2 disease.
  4. A nucleic acid sequence encoding the aminoacid sequence represented in Figure 2.
  5. An amino acid sequence which is coded by a nucleic acid sequence according to claims 1 to 4.
  - 20      6. An amino acid sequence according to claim 5, characterized in that it is a calcium dependent protease enzyme belonging to the calpain family, involved in the etiology of LGMD2.
  7. An aminoacid sequence according to claim 5 or 6 such as represented in Figure 2.
  - 25      8. An amino acid sequence according to claims 5 to 7, characterized in that the amino acid sequence is modified by deletion, insertion and/or replacement of one or more amino acids with the proviso that such aminoacid sequence has the calpain activity involved in LGMD2 disease.
  9. An amino acid sequence according to claim 6 to 8, characterized in that 30      LGMD2 is LGMD2A.
  10. A host cell unable to express a calpain enzyme activity, characterized in that it is transformed or transfected with a nucleic acid sequence comprising all or part of the nucleic acid sequence according to any one of claims 1 to 4.

11. Use of a nucleic acid according to one of claims 1 to 4 or a host cell according to claim 10 in the manufacturing of a drug for the prevention or the treatment of an LGMD2 disease.
12. Use of an amino acid sequence according to claims 5 to 8 in the manufacturing of a drug for the prevention or the treatment of an LGMD2 disease.  
5
13. Use according to claims 10 or 11, characterized in that LGMD2 is LGMD2A.
14. Use of an amino acid sequence according to claims 5 to 9 for the screening of the ligands of said amino acid sequence, said ligand being selected in a group consisting of substrate(s), co-factors or regulatory components.  
10
15. Use of a nucleic acid sequence according to one of claims 1 to 4 in a screening method for the determination of the components which may act on the regulation of gene expression of calpaïn.
16. Use of an host cell according to claim 10 in a screening method for the determination of components active on the expression of the calpaïn.  
15
17. A method for detecting of a predisposition to a LGMD2 disease in a family or a human being, such method comprising the steps of :
  - selecting one or more exons or their flanking sequences of the gene,
  - selecting primers specific for these exons, or their flanking sequences, or an hybrid thereof,
  - amplifying the nucleic acid sequences with these primers, the substrate for this amplification being the DNA of a human being; and
  - comparing the amplified sequence to the corresponding sequence derived from Figure 2 or Figure 8.  
20
18. The method according to claim 17, characterized in that the primers are those selected from the group of :
  - a) those described in Table 1;
  - b) those described in Table 3; and
  - c) those including the introns-exons junctions of Table 2;
  - d) those derived from the primers in a), b), or c).  
25
19. The method according to claim 17 or 18, characterized in that LGMD2 is LGMD2A.

20. A kit for the detection of a predisposition to LGMD2 by nucleic acid amplification characterized in that it comprises primers selected from the group of :

- 5        a) those described in Table 1;
- b) those described in Table 3; and
- c) those including the introns-exons junctions of Table 2;
- d) those derived from the primers in a), b) or c).

21. Packaging cell lines transfected by a recombinant vector, characterized in that the vector contains a nucleic acid sequence as claimed in claims 1 to 4.

10      22. Use of a packaging cell line according to claim 21 as a drug for gene therapy of an LGMD2 disease.

23. Pharmaceutical composition for the treatment of an LGMD2 disease characterized in that it contains a component selected from the group of :

- 15        a) a nucleic acid sequence according to claims 1 to 4,
- b) a cell line according to claim 24,
- c) an aminoacid sequence according to claims 5 to 9.

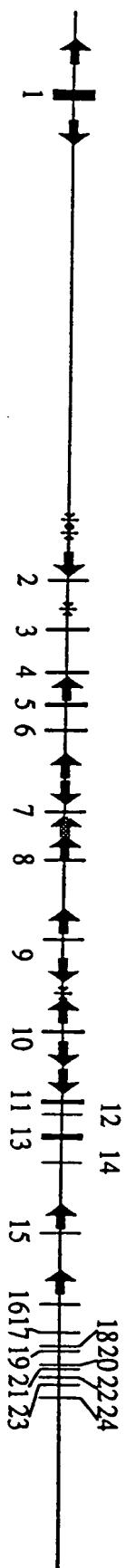
**ABSTRACT**

A nucleic acid sequence comprising :

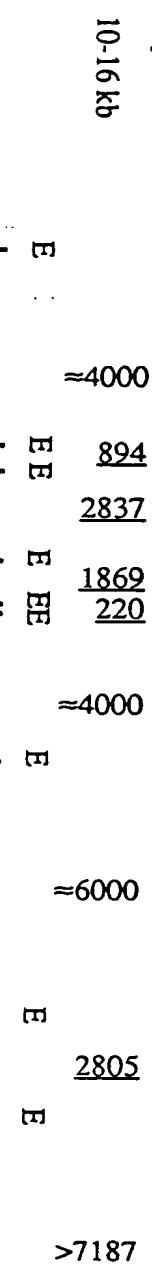
- 5        1) the sequence represented in Figure 8; or
- 2) the sequence represented in Figure 2; or
- 3) a part of the sequence of Figure 2 with the proviso that it is able to code for a protein having a calcium dependant protease activity involved in a LGMD2 disease ; or
- 10      4) a sequence derived from a sequence defined in 1), 2) or 3) by substitution, deletion or addition of one or more nucleotides with the proviso that said sequence still codes for said protease.

Figure 1:

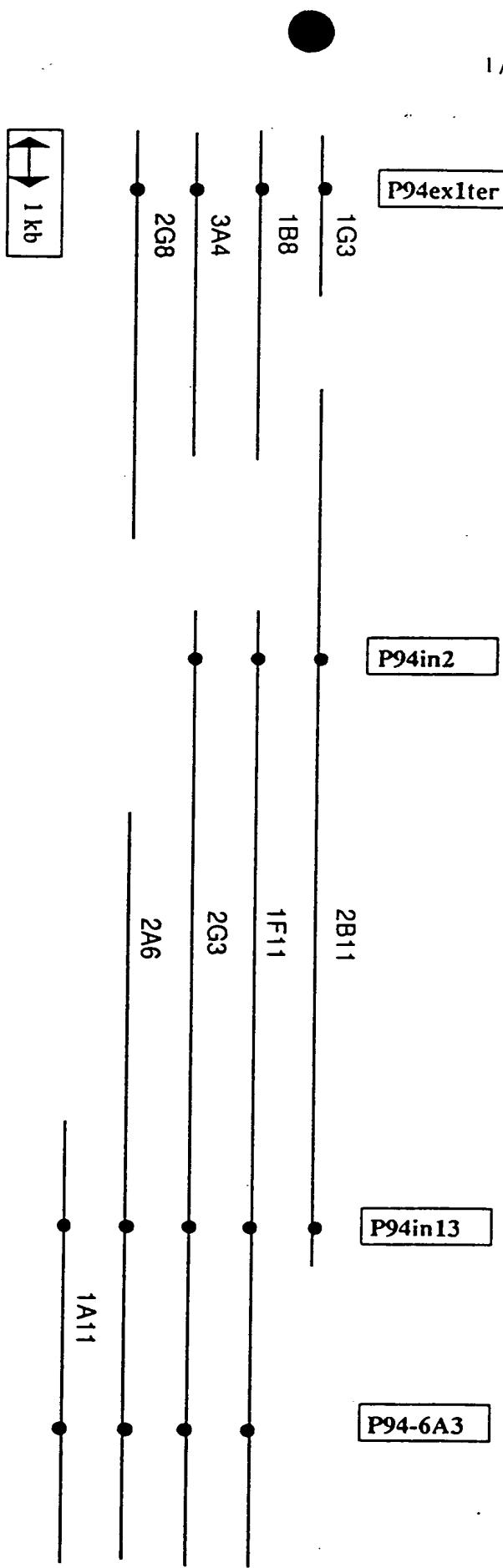
A) Genomic structure of the nCLL gene



B) EcoRI restriction map



C) Cosmid map

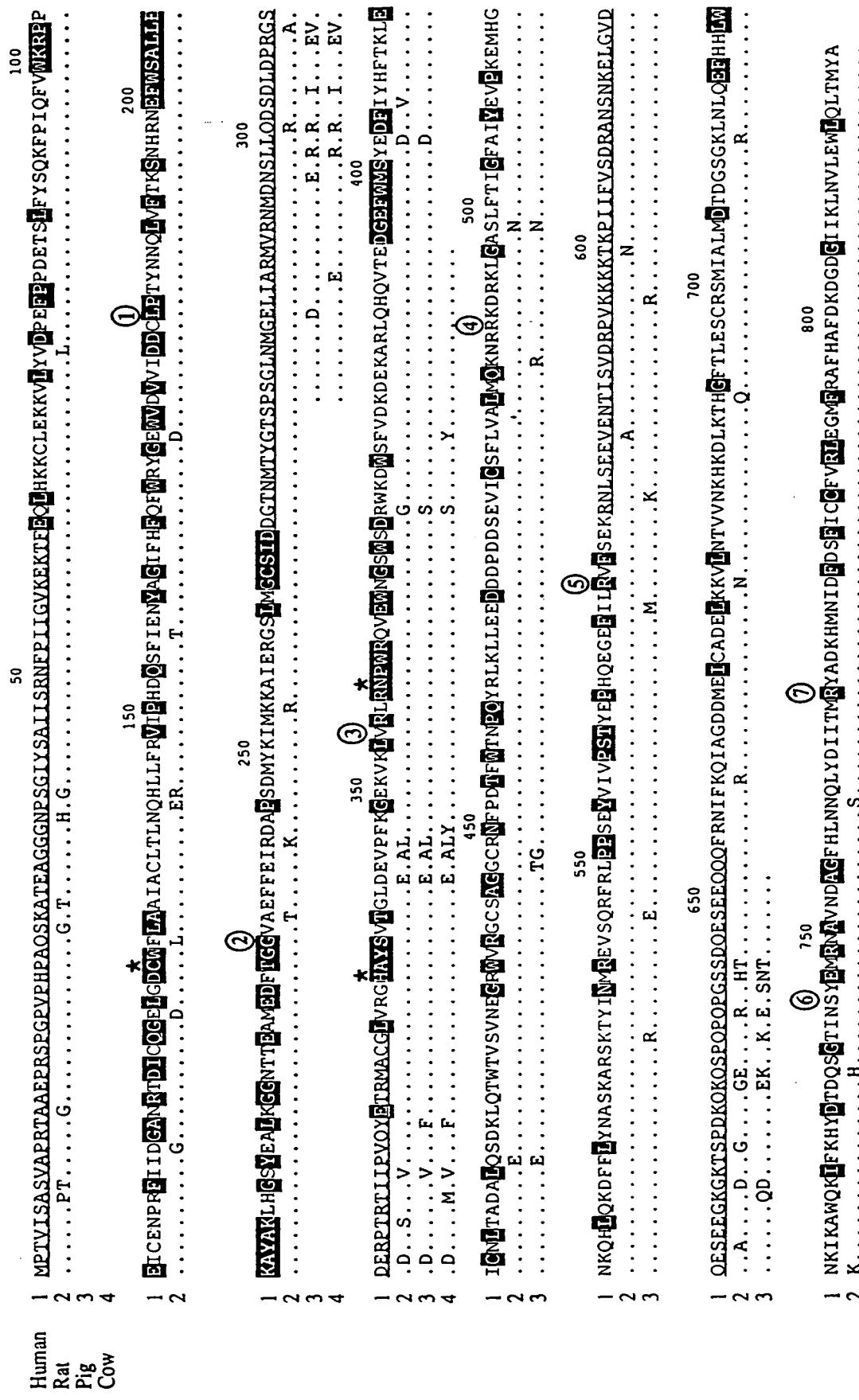


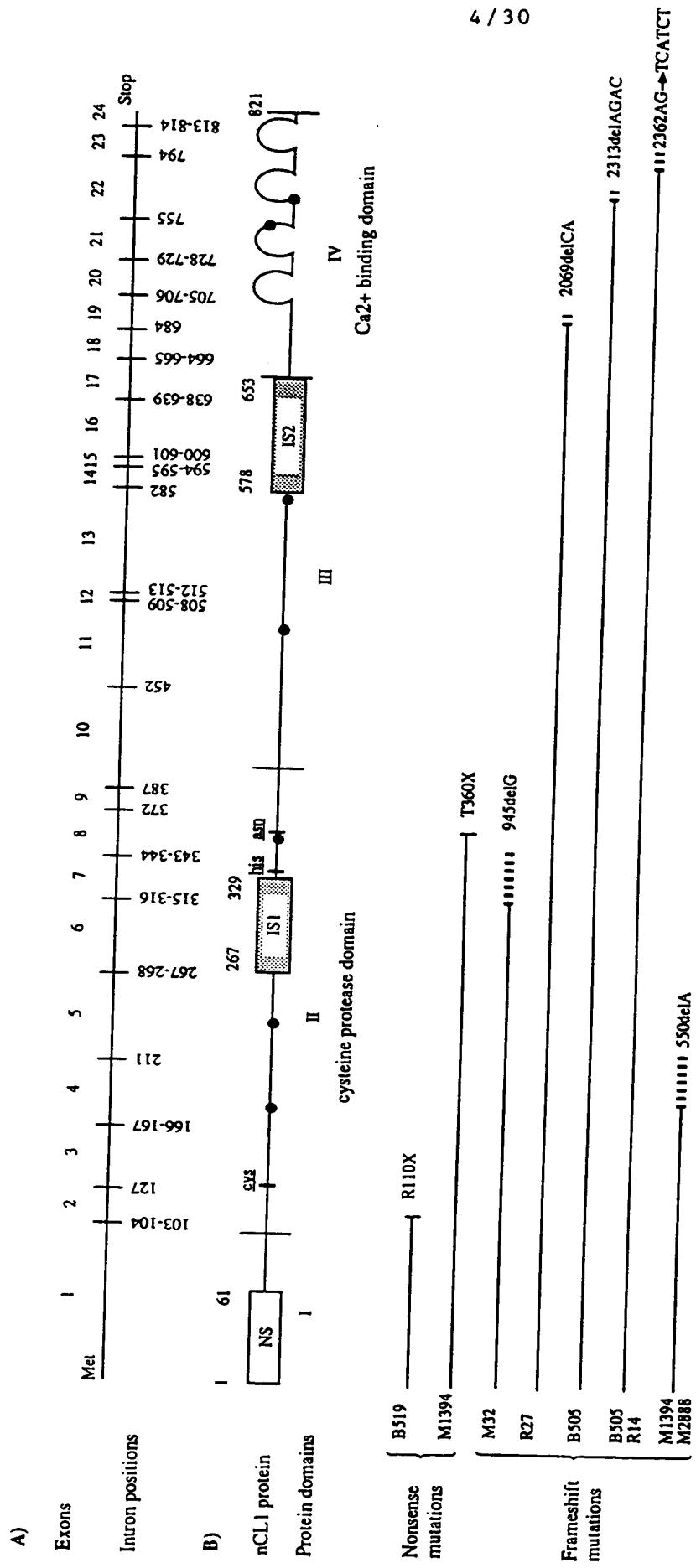
**Figure 2:**

A

6

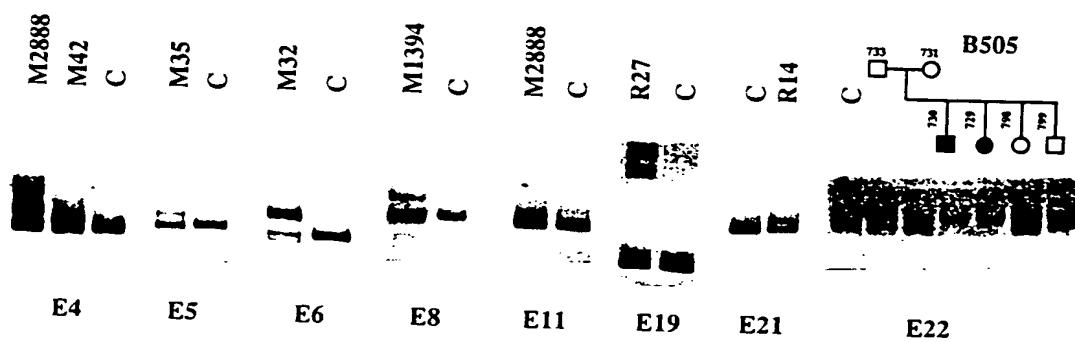
**Figure 3:**



**Figure 4:**

**FIGURE 5**

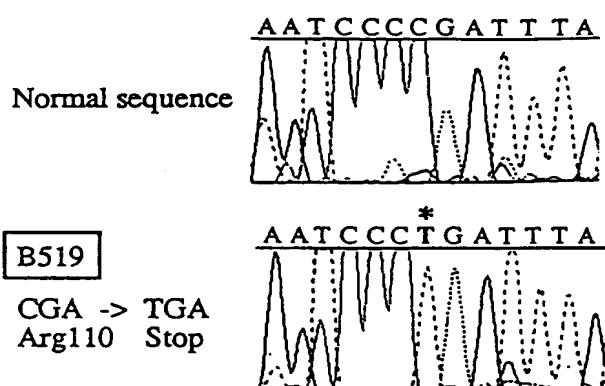


**FIGURE 6**

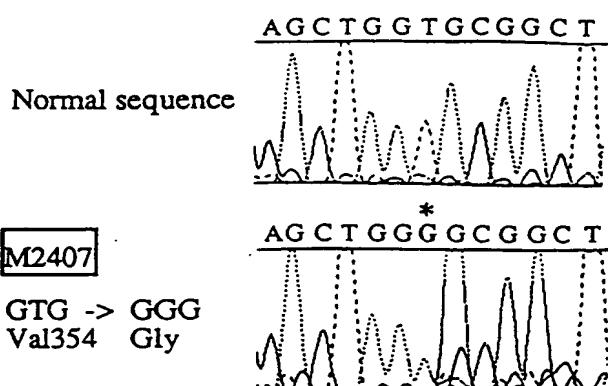
**Figure 7:**

7 / 30

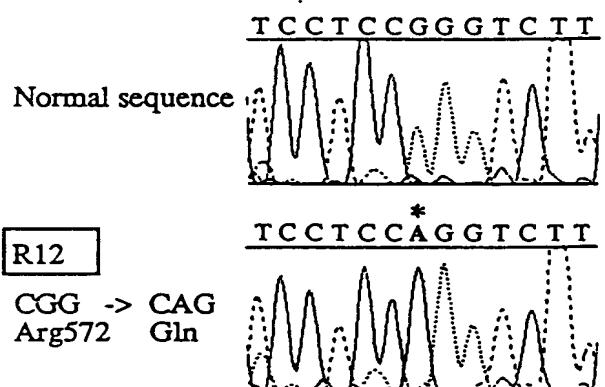
**A) EXON 2**



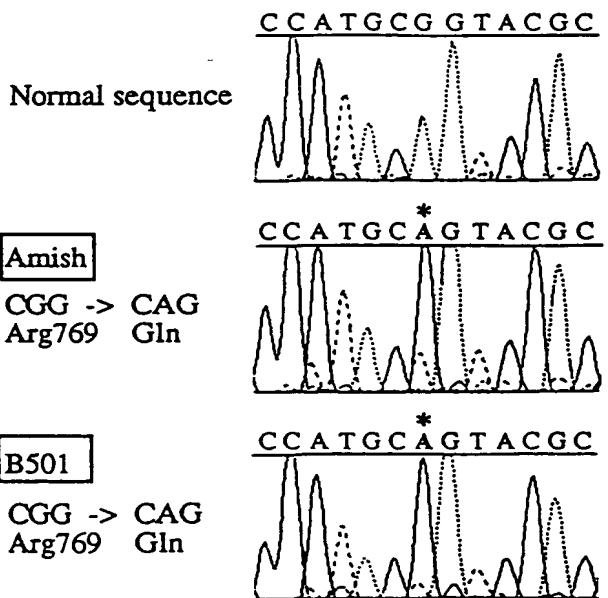
**B) EXON 8**



**C) EXON 13**



**D) EXON 22**



## LISTE DE SEQUENCES

## (1) INFORMATION GENERALE:

## (i) DEPOSANT:

- (A) NOM: AFM
- (B) RUE: 13, place de Rungis
- (C) VILLE: PARIS
- (E) PAYS: FRANCE
- (F) CODE POSTAL: 75013
- (G) TELEPHONE: (1) 45 65 13 00

(ii) TITRE DE L' INVENTION: LGMD GENE

(iii) NOMBRE DE SEQUENCES: 4

## (iv) FORME LISIBLE PAR ORDINATEUR:

- (A) TYPE DE SUPPORT: Floppy disk
- (B) ORDINATEUR: IBM PC compatible
- (C) SYSTEME D' EXPLOITATION: PC-DOS/MS-DOS
- (D) LOGICIEL: PatentIn Release #1.0, Version #1.25 (OEB)

## (2) INFORMATION POUR LA SEQ ID NO: 1:

## (i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 3018 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

(ii) TYPE DE MOLECULE: ADN (génomique)

(xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 1:

TGATAGGTGC TTGTAAACTG TGCTTAACGA AAACATACCG TGTGCTGTAG GGACTTAAC	60
CTTGGTTATA TCAGTTAGCC TGGTTTCGCT AACAGTACAT CATTGGCTT AAAGTCACAG	120
CTTACGAGAA CCTATCGATG ATCTTAAGTG AGGATTTCT CTGCTCAGGT GCACTTTTT	180
TTTTTTTAA GACGGAGTCT CTTTCTGTCA CCTGGGCTGG AGTGCAGTGG CGTGATCTGG	240
GTTCACACA ACCTCTGCCCT CCTGGGTTCA AGCAATTCTT CTGCTCAGC CTCCCAAGTA	300
GCTGGGATTA CAGGCACCCG CCGCCACACC CGGCTTATTT TTGTATTTT AGTAGAGACA	360
GGGTTTCACT ATTGTTGACC ATGCTGGTCT CGAAACTCGTG ACCTCATGTG ATCCACCCGC	420
CTCGGCCTCC CAAAGTGCAG AGATTAGAGA CGTGAGCCAC ATGGCCCAGC AGGACCACTT	480

TTTACCGAGAT TCAGTCCCAG TGTCATTCT GTGGATGGGG AGAGACAAGA GGTGCAAGGT	540
CAAGTGTGCA GGTAGAGACA GGGATTTCT CAAATGAGGA CTCTGCTGAG TAGCATTTC	600
CATGCAGACA TTTCCAATGA GCGCTGACCC AAGAACATTG TAAAAAGATA CCAAATCTAA	660
CATTGAATAA TGTTCTGATA TCCTAAAATT TTAGGACTAA AAATCATGTT CTCTAAAATT	720
CACAGAATAT TTTTGTAGAA TTCAGTACCT CCCGTTCAACC CTAACTAGCT TTTTGCAAT	780
ATTGTTTCC ATTCAATTGA TGGGCAGTAG TTGGGTGGTC TGTATAACTG CCTACTCAAT	840
AACATGTCAG CAGTTCTCAG CTTCTTCCA GTGTTCACCT TAATCAGATA CTCCCTTTC	900
ATTTCTGTC AACACCAGCA CTTCATGTCA ACAGAAATGT CCCTAGCCAG GTTCTCTCTC	960
TACCATGCAG TCTCTCTTGC TCTCATACTC ACAGTGTTC TTCACATCTA TTTTAGTTT	1020
TCCTGGCTCA AGCATCTCA GGCCACTGAA ACACAACCC CACTCTCTT CTCTCTCCCT	1080
CTGGCATGCA TGCTGCTGGT AGGAGACCCC CAAGTCAACA TTGCTTCAGA AATCCTTTAG	1140
CACTCATTTC TCAGGAGAAC TTATGGCTTC AGAATCACAG CTCGGTTTT AAGATGGACA	1200
TAACCTGTCC GACCTTCTGA TGGGCTTCA ACTTTGAACCT GGATGTGGAC ACTTTCTCT	1260
CAGATGACAG AATTACTCCA ACTTCCCCCT TGCACTTGCT TCCTTCCCT GAAGGTAGCT	1320
GTATCTTATT TTCTTTAAAA AGCTTTCTC TCCAAAGCCA CTTGCCATGC CGACCGTCAT	1380
TAGCGCATCT GTGGCTCCAA GGACAGCGGC TGAGCCCCGG TCCCCAGGGC CAGTTCCCTA	1440
CCCCGGCCAG AGCAAGGCCA CTGAGGCTGG GGGTGGAAAC CCAAGTGGCA TCTATTCA	1500
CATCATCAGC CGCAATTTC CTATTATCGG AGTAAAGAG AAGACATTG AGCAACTTCA	1560
CAAGAAATGT CTAGAAAAGA AAGTTCTTAA TGTGGACCCCT GAGTTCCAC CGGATGAGAC	1620
CTCTCTCTT TATAGCCAGA AGTTCCCCAT CCAGTTGTC TGCAAGAGAC TCCGGTGAGT	1680
AGCTTCTGC TTGCTGGCTG GTTTCCCCC CCACGGAGGA CTCTCTCAC TCAGCACCTC	1740
CGGCAGCTCA GCTGTGCACA TGGGCACTGG GGGAAAGGATC CTGGCAGCAG CTCTGCTGGG	1800
CTCTGTCTT AAGTGTGAAG CAGGGAGGAG AGGAACAGGT CTCAGATATT TCACCAAATC	1860
TCAGCAAAAT CCAGAGGGAG AGCCGAGGAG GTGGGGTGAT TCTTATGCTC TGGCTTTTC	1920
TCTCTGAAAA AAAAAAAAT ATCTTGCTTT TTATAAAAGT GGGTGGAACT CAGTTAATT	1980
CATCCTGTAA AAATAAATAT TCCTTCTCA GAACAAATTG CAGACAGCCC AGATGTACCT	2040
GTTCGTTTA ATATTATTCA TCTTGGTAAG ATTATTCAAG TTTCTCTGGC TAAAATCATG	2100

ATGTTATTCT TCTTTAATT ACCAATGGCC ATTCTTCTG AAACACAGAA ACCCTAGAAA	2160
GAGAAGAGTC ATAGGCAAGG AATTTTTTC ATGCATAAAA TGTTGGGTT AAAGAGAGAG	2220
AGACCTAGCA ATCGCTTGG TCCACCTACC TCACCTCATA AGTGAGGAGT CAAGGCACAC	2280
TAGAGTGAAA TATATCTAGT GGGCACATGA CAGAGCCCGG ATTAAAACCTT TGTTTAGGA	2340
AACTCTCCA GCCTCTGGT TTCATTTACA GTGATGCCA GGAGGGAAAT CACATTCCCC	2400
TGGCTCACCT CTCTGATCAT CCCTCCAGTG TGACTCTTGT TCTTAATTCG AGAAATATTT	2460
ATTGAGCATE TACTAGTGCC AGCACTGGGC AAGCAACTGG GGGGACACCA GTGAGTAAGA	2520
AAGACCAAAA TTCCAGCTGT CTTGGAACCT AGGGTCCTGA AGGGAAGATG GGCATTGAAC	2580
AAGAGTGACA TTGTCAGGAG ACGATGTTCT GGGTGCCACA GGATCATGTG GCAAGGAGAG	2640
CTAACCTGGT CCAGGGAGAC AAACCCCTCTC TGAGGAAATG ATGACAAGCT GAGACCAAT	2700
ACTATTGATT AGCCATGGTT TTCTTTAACC TAAGGTGGGC CAGGCATGGT GGCTCATGCC	2760
TATAAACCCA GCATTTGGA AGGCCAGGC TGGAGGATTG CTTGAGCCCA AGAGTTAGAG	2820
ACCAGCCTGG GCAACAGGGT GAAAACCTAT CTCTTTGTA CTAAAAATTTC AAAAATTAT	2880
CCAGGCATGG TGGCACATGC CTGTGGTCCT AGCTACTCAG AGGCTGAGGT GGGAAAGATCA	2940
CTTGAACTCG GGGAGTTGA GGCAGCAGTG AGCCGAGATC ATGCCACTGC ACTCCAGGCT	3000
GGGTGACAGG AGTGAGAC	3018

## (2) INFORMATION POUR LA SEQ ID NO: 2:

## (i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 11451 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

## (ii) TYPE DE MOLECULE: ADN (génomique)

## (xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 2:

GATCCACCCG CCTTGGCCTC CCAAAGTGCT GAGATTACAG GTGTGAGCCA CCACGCCAG	60
CCGACACTGC CCTAACTCTC AAGTTGCATC CTTACTCGAA TAGTATGACA GTGTGGAAAG	120
CAGCATGGGA CAATGTAAAA AGGAGGCATG TTTCTGGCTT CTGCTACTTA CTAGCTGTGT	180
GTCTTGCAC GAGTTCTTA ACCTCTCTGG GCCTCAGTTT CCTTATCTGA AAAATAACAA	240
TGATAGTATT CCCTTCACAG GCCCAAATGG AATACTATCA GGAACACTAC ATAATGGAAC	300
TCAATAAATA ATAGCTACTG CGGCCGGGCG CGGTGGCTCA CATCTGTAAT CCCAGCACTT	360
TGGGAGGCCG AGGCGGGTGG ATCACAAAGGT CAAGAGATGG AGACCACCT GCACAAACATG	420
GTGAAACCGT ATCTCTACTA AAGATACAAA ATTAGCTGG GCATGGTGGC GCATGCCAT	480
AGTCCCAGCT ACTCGAGAGG CTGAGGCAGG AGAACACTT GAACCCCGGA GGCAAGAGTT	540
TCAGTGAGCC AAGATTGCAC CAGTGCACTG CAGCCTGGCG ACAGAGTGAG ACTCCGTCTC	600
AAAAAAATAC CTATCTATCT ATCTGTCTAT CTACTGTTAT TCTTACCTGG TCATTTCTT	660
TTTGTTCAC AGGAAATTG CGAGAATCCC CGATTATCA TTGATGGAGC AACAGAACT	720
GACATCTGTC AAGGAGAGCT AGGTAGGAAA GTGCCTCAGG TCAGATCCTG CCAGATGATC	780
AAGGGGTGAT TACAAGGTGT GATCCCCTTC CAGGAGGTAA AGGGACAATC TGTGCTTGCT	840
TCCAGTAACT TTTTGGAAAGA TTTTTATAA CAGTTGCTTT ATGGTCGTTT ATCTACATGC	900
TGGCGATTGC TTCATTCCT CCTACATGCC TCTTAGCAC TCTGCCATGC ATCACAGGG	960
GTATCTGCAT CCTGTGGCCT CCTCTCCAGT ATCTCAAGGA CACTTACATA CCCCCACTCAG	1020
CATGACAAAAA CCCCTGCTTT TCACGTATC GTCTTCTTG GAAGACAGCT CTGTGACTGT	1080
GCACCAAGCA TGCCCCTTGG GCATGGAGAT TCTAGATACA CACACAAAAG GCATGCCAA	1140
GGAAAGCACT TGTAACCTGGA ACCCTTGGTT TAAATTGGCC CAGCATAGCT CCATTTAA	1200

AAGAGTCTTT CCACAAAGAT GGCATCCGCC ATGTGGATGA GCATCCAATT TTCTCTTG	1260
TTGGTTAGCT TGACTGCTCC ATCTGATCTT CCTCTCTCTC GACCTCTTGT TCAGAAAGTA	1320
TTGTCTTGG TGTGGACTAT AAGCAAGCTC TGTGAAGTAA AATTGGAGAG AACACCAACA	1380
GAAACAATT AAATTGAGG AAAAGGGGGC ACCTAAGACC AAAGGAATT GGCTTATTTC	1440
ATTCCAGAAG GGGAGGCTGA GAATAAAATCA GATGAATATC TGGGTTCCCTG CACCTGAGGG	1500
AAGGCTTCCT GCAGAGCCCT GGGCATAATA ATCTGGGACC TTCAAACCAA TAACCTCTTT	1560
TCCAAGGAAA GACTGGCTGC TTCCAAGGAG GGTAGGGAG AGTCGGGCTG CAGGCAGCTC	1620
TCAAGTCTCC CCTTGCACAC TCTCAGGTTG GCATTTCAC TTTAACCCAT CCTCCCTTAA	1680
GAAGGCAGTT CTTTGTGACC ACGGTACACC CCCTATTATA TATATATATA CACACACAGA	1740
GAGAGAGAGA GAGAGAGAGA GAGAGAAAGA GAGCAAAGTG TTACCTCCAA CTACATACAG	1800
TACTCTGTCA GAAAAGAGGT TCAGAGAATA AGAAAACGTC CCGAGCTCAT TCCGTTGCCA	1860
GCAATGTCTT ACTGCCCTCT ATAGACGGGT TCCAGGGCAG CTGCCTACCT GGCCTTCCTT	1920
CCAATACAAA TCATCTTGGT GGATGGTTCT CTGAGGCTCA GTCTTCGCTG AAGTCAGAAG	1980
AGGAATTGGA CTCACATTGC AAAGGCACAG GCCAGGGCAG ATTTCTACA GGTGTTAGGA	2040
AGAACAAACCC AGTTATGATC ACCTACTGCT CTGTCTCCAT TGAGGCCTAA AAAGGAAGTG	2100
AGTTTATACT GCAGTTGGAG GAACTGCCTG CAGCCTTGAG GAAAATGTCT AGTCACAAGG	2160
GAGTAAGTTA CCTGTTGATC ATATTGTCAA GGAATTCTG TCCAATTCTC CTTCCCTGGG	2220
TTGACACCTC TGTAAGGTCA GATCTGGAAG TAGGAGAGTG GGCACCAAGG GAGTCCCCGT	2280
TCAGGGAAAGT GGAGTGGCTG GCTGGGATTG GGGCTTTTC TTCCCAGGAG GAGCAGGAGT	2340
GCTCACGATC TGTGCCCTGT GTCTGCCTGC AGGGACTGC TGGTTCTCG CAGCCATTGC	2400
CTGCCTGACC CTGAACCAGC ACCTTCTTT CCGAGTCATA CCCCATGATC AAAGTTTCAT	2460
CGAAAATAC GCAGGGATCT TCCACTTCCA GGTGAGGTAA TGAGAGTGTA GTTAAGAGGG	2520
CCAGCGGCAG GCCACCCACC GCTGGTCTCC TGGCCTTGAC TTCCCAGAAG CTGGAGGAAA	2580
CTTCCCACCC ATCTACCCGC AGCGGCAACA GTCGGCATGG ACCCCCTTAA GGCTTCAAGC	2640
CTGGGAGGAA GCAGTTGCTT ATCTCTGGCT CCCTAATCCC TCCCCCACCA CCTTCCACTA	2700
TGTCCCAGAA AGACAGGAAG ACATCCTGTT TACTGTGGGT CTATTTTGT CTTTGCAGCT	2760
GTCTGGCTGC TTTTATTGCC TGCAGCCCTT CTCAAGTAGG TCCCTAAGAT ATTACCACTG	2820

TGACACCACA GGACCCTCA GGTTGTACAG GAACCCCTGT CCAGGGCTCC TGTATACTTC	2880
TTCCTCTCTA AGGCATGGCG GTACCAAGGC TATCACTCCT CTCTTCCAAG CCCTGGAAGA	2940
AGAGTCTGCT TAACCTGGGG ATCAGGCTTC TTGTTGCC TAGAACTGAA TCTGATGGTT	3000
CTAGAACATCCA TCCAGCTACT GGAAATTTC TGGGTCCCAG TCACCTTGGC ATAGAGCTGG	3060
TGCTAGAGCA GAACCAAACG GAATTCTACC TGTGAGGGTC TCGTAGCTTC CGGGATGCTG	3120
GGGAGTCAGC CTGTCCTCAG CTTCAAAGGC TCCCTCATGT CCCAGGATGA CCCACATTAT	3180
CAGTTCTTGC TCCCCGGGTC TTGCACCTCA GCACGGAAGG CCTCAGAAAA GGTCTGTCTC	3240
CAGGCTCAGA CTCCCCCTCC TGCCGCCCTTG GGAACATGGC ATATTTAAAG GGTCTCAGAT	3300
CTAAAGGGCC TTACATACAA ATATCAGATA GATTCTGTT CTCATTTCAA TGAGGGAGAA	3360
AGTGCCATTG AAAAGGAGAC TAAACCACAT TTGGCCCTTT TCAGTTCAAA CTGATTCAATT	3420
CAAAAAAGAG CGACATCCAA ACTTGAAATG ATTGAACAAT GTTCTGCTA CAGCTAGAAT	3480
AGATTCTGGG TCACTTTGTT CCTCCGTTTC AATCCTGTT CTTCAGTTTG GCATCAAGAA	3540
ATACCTAAAT CAGCACAGTG CCTTCACTGC ATAGTTCCAAT ATCCTGGCCA CATTGAATCA	3600
GCTGGGGCA CCTGAGAGTG CTGACACCCA GGCCCTGCCAG CAGACCTGCT GAGCAGGAGA	3660
ATGAAAATCT TACATCCTAA GACACTCATG GAGCACCTAC TCTACCCATT ACTGGGCTGG	3720
ACTCTGTGGA AGACATGAAG TATATGTAAC TCACTTCCAG CTCTCAAAAA GCACCCAGTC	3780
CAGTTAGAGA CAGATTACA CACCCCAAAC ACAAAATAGG ATGAACAGGC ACCCAGATGC	3840
AGAGTCCAGG AAATGATGCT GCTTGGGAT TCAAGAACCC CCTGAGGAAT GTGGAGGAAG	3900
GACACATTTC CTAACAGTAA TTTGAGTATG TGACTCTGTG CGTGACGCTT CTGTGGAGTT	3960
CTGGCGCTAT GGAGAGTGGG TGGACGTGGT TATAGATGAC TGCCTGCCAA CGTACAACAA	4020
TCAACTGGTT TTCACCAAGT CCAACCACCG CAATGAGTTTC TGGAGTGCTC TGCTGGAGAA	4080
GGCTTATGCT AAGTAAGCAA CACTTAGAA TGTGAGGTGG GGCTAGAGGT GAGAAAGTGG	4140
GTTGCAAAAT CCAGCCGAGA CCTCACTCAC AGGAAGAGGC ATGTGCCTCT ATACGTGCAT	4200
ATGTGTGGGC ATGCAAGTCC AACTGTGACC CAAAGTTAGA GATCAGTTCC AGGCAACAAAC	4260
AGCTCTAACT AAAAACATTA AATTAAAGAG TAGAAATGAA GATTGCTATA GAAGACCTTT	4320
AGCTTTAGCT CACCATAGCG AGTTCTTCA TTGCACCTCC ATGGTGGCAT TGCAAGTCTT	4380
GGGATCAGAG CATTGTCCCA GGGTCTCGAT TGGCTCAACC TCATGTGCTT ATAGAAGATT	4440

TATAAAGACA	TGTTGTCTCT	CAACTAAAAA	GCTCCACCCC	AGATGATAAT	AATGGATT	4500
CAAATTTTGG	AACAAGGTCA	CTCTGTAATG	CAGGCTGGAG	TGCAGTGCTG	CAGTCACGGA	4560
TCACTGTAGA	TTGACCTCCT	GGGTTCAAGG	TGCTCCTCCC	ACCTCAGCCT	CCCAAGTAGC	4620
TGGGACTACA	TGCGGGCATC	ACCATGGCCC	TTTTATTTTT	GTATTTTTTT	GTAGAGCGGG	4680
GTTTTCCCAT	GTTGACCCAG	ACTGTTCTCG	AACTCTGGG	CTCATACAAT	CCACCAGCCT	4740
TGCCCTCCCG	AAGCGCTGGG	ATTGCCGGTG	TGAGCCACCA	CACCGGCAGC	TGCTAATGGC	4800
TTTAATGCAG	CCCTTCCTCA	ACGTTCAGGA	TGTAGTGGAA	AGAGCTCTCA	GGAAGTGGGG	4860
ATAGCTGGGT	TTCAATCCA	GTGCTTCTGG	CTCTCTGTGG	TCTTGGGTGG	GTCACCTAGC	4920
CTCTTGAGCT	CAGTTCTTC	ATTATGAAGA	AAGGGAATCA	TTGTTCCAT	CCCATGAGCT	4980
CATAGGGTTA	ATGTGGAATT	GATGAAAGAA	CATCACAGCA	TCCAAGAGGT	AAAGTTCTGG	5040
TGGCAGTGGT	ACCTGGTTT	TGTTCCCTGG	AACTCTGTGA	CCCCAAATTG	GTCTTCATCC	5100
TCTCTCTAAG	GCTCCATGGT	TCCTACGAAG	CTCTGAAAGG	TGGGAACACC	ACAGAGGCCA	5160
TGGAGGACTT	CACAGGAGGG	GTGGCAGAGT	TTTTTGAGAT	CAGGGATGCT	CCTAGTGACA	5220
TGTACAAGAT	CATGAAGAAA	GCCATCGAGA	GAGGCTCCCT	CATGGGCTGC	TCCATTGATG	5280
TAAGTCTGGG	GTGTGGGCA	CAGGGTGGGG	AGCTCCAAGT	GTCAGGAAGC	CTTTTACCCA	5340
ATGAAGGGCA	GCATAGAGCT	TTTGTGTGGG	ACAGAGCGAA	TGTTTTGTTT	GAGGAAGCAG	5400
GAACCTGGCTC	TCAACTTTGA	GGACTGGAA	TTTCTCAAGG	GAGAACAGTT	CTTCCGGATT	5460
TTCAATAAAG	ACACTGGTCA	AGGACATTTC	AAGCCCTGGA	ATGTCAGTGG	AAATCAGTCC	5520
AGAGGCCTGT	GTCAGTGGAG	GCCTCCCTTG	CTGGTGCTCC	TCAGTCTCAG	CACGCTCCCA	5580
TTAAGCTGGC	CACGTACTTG	GCTGTGGACC	TGAGCCCACC	ATTTCCTAA	GAAAGCCTCC	5640
CAGTCACTGG	GCTTTCACCA	CACCTCCCCG	CTTGAGACGT	GGGTTTG TG	TTGTTACCTG	5700
GGAGAAGCTA	AGCCTGCAGC	ACCTTTCAGT	GCAAAGAAAT	GCTGTGAACT	GAGACAGGAG	5760
CCAACGGTAG	GGAGATGGCC	GCCCATGGCC	AGGCCTCCTT	CAGGGGGCAT	GCCTTCCCTG	5820
AGGGCTGCTC	AGTATATTGA	TATGATAATC	TTAGTGGTTT	CCATTGGGA	GGATGGGCT	5880
GAAGCTGAAT	TCCTGCCCT	TCTTCTCCC	ACACGCCAA	TGGACAGCTT	CGAAGGTCA	5940
TTAGCACACA	ACACCATGGA	TGAACTTTTT	TTCTGTATCA	CTTTCTCCG	TCTTCCCTCC	6000
ATTCTGCTC	TGTTGATCTC	TCCTCTCTCC	CTTTGTCTGT	CCCATCTCTT	TCTCCTCTCT	6060

CCTTCCCTTT CCACCCCTCT GTGTTGTT CTCCTCCC CTGTGTTGTT CCCTACATTC	6120
TCCATCGGGC CTCAGGATGG CACGAACATG ACCTATGGAA CCTCTCCTTC TGGTCTGAAC	6180
ATGGGGGAGT TGATTGCACG GATGGTAAGG AATATGGATA ACTCACTGCT CCAGGACTCA	6240
GACCTCGACC CCAGAGGCTC AGATGAAAGA CCGACCCGGG TGTGTACACC TCCGATTATC	6300
AGAACTGACC ATCCCTCCAA CCCACATGAC CCCGCCCTAT TAGTGTAGA CTCCCCTCAG	6360
CAGCCAGGGC CTTACCCACA CACCCCCACC TGGCACCTCC CAAGGGTCTG GGTTGAAATA	6420
ACTTGCTCAG CCAAGGCTCC TGAAGAGGGT GCAAGAACCA GGATTTGGA GGGAACTCT	6480
GCTGGAGTTT CTGCATATT CATGGTCCAG GCAGTTCTC TCATAACGAA CTATCAGACA	6540
GAAATACTTG TAAAGATACT TCATTTATTT TGAAATATTT TTCTCTTCT AATGTATTCA	6600
TTTATTCACTT CAACACTTAT TTTGAGCTC CTACTATGTT CCAGGCACTC CTCTAGCAA	6660
CAAAGCAAAT TCTCTCCTCT TTTCAATAT TTGTGGAAAA AGCAAGGTCT CCCTCTTGTA	6720
GAGTTTATAT TCTAGTATTTC TCATAAGTTA TACCTGCTCA CTGGAGAATA CTGAGCCATA	6780
CAGAAAAACA CAGAGGAAAA TTTCACTTAT ATTTTCCCC ATGTAAGAT AACCACTCTT	6840
AACATCTAGT ATATGTTCTT CCAGGATTT TCTATGCACA CACTGAATCT GTATTTTAT	6900
TTTTAAAATG TTATCATATT GTATGTACCT CTTTGCAGCC TGCTTTTTC AGTTAGTTT	6960
TTTGGTTTTT TGTTTTTTT TTTTTTTGG AAACCAAGTC TTGCTCTATT CCCTAGGCTG	7020
GAGCACAGTT GTGCCATCT CGGCTCACTG CAACCTCTGC CTCCAAAGTT AACTAAATT	7080
TCCTGCCTCA GCCTCCCGAC ATAGCTGGGA TTACAGGCAC ACACCACCAC ACATGGCTAA	7140
TTTTGTATT TTTTAGTACA GACGGGGTTT CACCATGTTG GCTGGAATGG TCTTGAACTC	7200
CTGACCTCAA GTGATCCACC TGCCTCAGCC TCCCAAAGTG CTGGGATTAC AAGTGTAAAGC	7260
CACCAACACCC GGCCTAGTTT GATATTCTTA ATGTGCCAA AGTATTCTCC TGTAACATT	7320
TTTAATAGCT ACACAATATT CAAACACACA GATATGTTAT AATTTATTTA CCCAATACCC	7380
TATTATTGGA AAGTTGAGTT CTTTTTTTC TTTGTTTGT TTTGTTTGG TACTATTCTA	7440
AAATGCTATA ACGAACATCC CAATAGATAC ATCTTGTAT ACATCCATGG TGACTTCCAT	7500
AGGACAGATT CCCAGCAGTA GAATTGCTGG GTTGAATGAT ATGCTTAGGG TAATGACAGA	7560
AGACTCATT CAAGCAGCTT CCTAGGGTCT TAGAACTTAA GGATTAATGA GTCTTCCCGC	7620
CCCCCTCCAG TCTATTCAAGC ATGATCTGGA TCATGAGGAC TGAGATCTGG AAGAGACTGA	7680

GATCTGGGAG AGGCTGAGAT ACCAAAAGCC CTGGCTCCAC CCATAACCCCT CGCCCTGAAA	7740
ACAGCTCTAG GAATTCCGGG GCCTAGCAAG GCTCCGGAA GCTCCTTTA AAGCTGTGAC	7800
GTTAGTAGGC ACATGGACCA TAGAGACCTA TCCAGGGCTC ATGGGACTTT AGTGATCCTG	7860
CCCTTCTCCC AAGGATCCCC CATGGCTGCA ACTTGGAAAT TTCTGCAAAT GGAAGAGCTA	7920
CTCCTTAGGC ACGGTCATGT CTGAGCAGGG ATCTCCTCGG GCTTTCTTAG AATTCTCTCC	7980
CTGGGCAC TG GGACTCTTGA TTTCTTGAAT ATTATGTTCC AGGTGGGTGT GGAGGAGGTG	8040
AGGGGATGTA AAGAACGGCTA GACTTGGCCA GGCGCAGTGG CTCATGCCTG TAATCCCAGC	8100
ACTTTGGGAG GCTGAGGCCG GTGGATCACC TGAGGTCAAG AGTTGGAGAC CAGCCTGGCT	8160
AACATGGTGA AACCCCGTTT CTACTAAAAA TACAAAAAAAT TAGCTGAGCA TGGTGGCACG	8220
TGCCTGTAAT CCCAGCTACT CGGGAGGCTG AGGCAGGAGT ATCGCTGGAA CACGGGAGGC	8280
AGAGATTGCA GTGACCCGAG ATCGCGCCAC TGCACTCCAG CCTGGGCGAC ACAGCAAGAC	8340
TCTGTCTCAA AAAACAAAAA AGAAAGAAAA AAAGGAAAAG CTAAGACTTA CATGTGTCAC	8400
TTAACCCCTT TTCTCAAACC TCTTCTCTT CCAGGAATAG TCAACCCCTG GATGGCTTCA	8460
GGGGAAGGGG GATCCTGAAG CCCAGGGCAG CCTCCAACTC TACCCCTTCC TCCTTGAAG	8520
GATACTAAGG GGTCCAGAAA GGAGGGGCAG GACACTGTTA CCCACCCAC ATCCCAGCAT	8580
CCACATTGCT CTCTGATGGT CAGGACAGAG CCTTCTCAGG GAGACCAGCC TGTCTGGAGC	8640
TGTGTCTCTT GGCACTCTTA AAGGGCCACT GAAGGTCCTG TCGTGGTCGT GAGGCACACT	8700
TTCAGGGAGC AGAGTGGTCT GTGTCTTCAC AGAGCCCGGA AAATGAACCA GTATGAACCTT	8760
TGCCTCCAAG CAGCAGAACT TCTGTTCCCC CGCCCCTAAT GGTTCTCTG GTTACTGCTC	8820
TACAGACAAT CATTCCGGTT CAGTATGAGA CAAGAATGGC CTGGGGCTG GTCAGAGGTC	8880
ACGCCTACTC TGTCACGGGG CTGGATGAGG TAAGCCTGGT GGGGCTTGGT GGGGCAAGGG	8940
CACCCCTCTG GGTAAACCTC ATGAAGTCAG GACTTAGCTG TTGGGGCCCC TGCCCTGTCT	9000
GCAGAGCTTG CCTCCAATCA GGACATTCAAG TTCAAGGTCC AAGCCACGCC TGGGAGCAGA	9060
GGGGCCTGTG AAACCTGGTAG AGGTGGATCC TGCCACAGTT GGTGCACAGT TTATCTTGC	9120
TTTCGTGCT AAAGATGGCA ATTTTCCAA CATTTCATG GAACAAATTG AAATATCACT	9180
TAACTTGCT TTTACAAAGT TGTTCTGAG TGTTCTGAG CTTCTGTTC TCTCGTGTTC	9240
AGATAGCTAC AGTTGTCTCT GGGTAGCCAC GGGGACTGGT TCCAGAAGCC CCAACAGTAA	9300

CAAAATCTGC AGATGCTCAA GTCCCTTCTG TAAAATGGAG TAGTATTTGC ATATAACCTA	9360
TGCACATCCT CCCATATACT TTAAGTCATC TCTGGATTAC TTACGATACC TAACACAATG	9420
GAAATGCTAT GTAAATAGTT ATTGCACTGC ATTGGGTTTT TTTGGTATTA TTTCTGTTG	9480
TTGTATTATT ATTTTTCTT TTTTGATAA TTTTGATCC ACAATTGGTT ATATGCCAAA	9540
GCCATGGATA CGAGAGGCTG ACTGTTCTGT TTTGCTCCTT CTGGGACTTC TGGGTTTCC	9600
TGGACCATGT CTGAGACAGG AACGTTGTAAC GACCTGTTGC ACACAGTTGG GCAGGTTGTG	9660
CCCTGTACAG AGGGATGGGC TGAGAGGGGC AGTTGCCTGC ATCACCCATT GCAGCAGACT	9720
GGAGGGAGTC TGCTTGTGTTG TAGTTCCCTCA GTCAGCAGGG GCCTTTGTC TTTCCCTCCT	9780
TTCCCTTTTT TTTTTTTTG AGACGGAGTC TCACTCTGTT GCCCAGGCTG GAGTGTAGTG	9840
GCACAGTCTC GGCTCACTGC AATGTCCGCC TCCTGGATTCA AAGCGATTTT CCTGCCTCAG	9900
CCTCCTGAGT AGCTGGGATT ACAGGCGCGT GTCACCATGC CCAGCTAATT TTTGTATTTT	9960
TAGTAGAGAT GGGGGTTTCT CCATGTTGAT CAGGCTGGTC TCGAACTCCT GACCTCGTGA	10020
TCCGCCACC TCGGCCTCTC AAAGTGCTGG GATTACAGGC GTGAGGCCACC ACGCCTGGCC	10080
AGCAGGGGCC TTTTTCTAA TTTATATGAA GACACCTAAT TTATATGTGT TAGCAAAGCC	10140
CTCCTGTTA TGCCTCACCT CCTCCCCCGA AGCTCATACG GCAGGATGTT CCTGAGAAAA	10200
TTGCCTCTTA GAAGATAGAG AGGAGATGCC AAGCCTAAGT TAGGCAGACT CAGGAGGATA	10260
GGTCTGACCC ACCCCCTGCC ATTCCCCAGC ACACTTGTGA TTAATCTCCT TGGCCAGAGC	10320
CAGGCAGAAC ACCCTCCCGT AAGAGATTTG CCCCCCAGCC CCGTCCCAGC CCTCAGCTAG	10380
ACAGAAGATT CCCTTCCAG AGAGGCTGCA GAGCATGAGA GCTCTTCTG TGTGCTTAAG	10440
GTCCCGTTCA AAGGTGAGAA AGTGAAGCTG GTGCGGCTGC GGAATCCGTG GGGCCAGGTG	10500
GAGTGGAACCG GTTCTGGAG TGATAGGTAG GTGAGGGGAC CCCACGGGAT TGGCGTGGC	10560
GGGAAACAGG GTCCGGGACA AGGCTGTGTT GGGAACTGAG CCATGAGAGT ATTGAAGATG	10620
CTTGGTATAA AATCACCCCTC AAAACCAATG ATCCGCAGAG AAGAGGGCA CAGGTGTTGG	10680
CTCCAGGGAA GGGCCAGGAG TGGAAAGCGGG GTGCTGGGA CCCAGAGAGG TTGCTGACAA	10740
CCATTGGCTG GAAAGGAAGG ATTCCAGAAA GCGTGGGAA GGTCCAGGCA CGAAAACCGT	10800
ATGAATGCAG GGTTCTGGC TAGAGAAGTG ACTTCCCTTC TTGGGGTCTT GTGTTGCCTT	10860
TCCTGTGAAA TGGGAACAGT ATTATTAGCA CTTACCTTGT GGGCTGATAT TGAGGACTAA	10920

CTGGGACTTG TTTTGGCCA AGTGCTGAGC CATTGCTAAG ATTCCCTTA CCCGTGCTTG	10980
TCCCTTGTAT TAAGGCACAA GGGCCCTTC AAAAGAATT TACCTGCTTT ATCAATTGAA	11040
AGGGATTAAG ACCTTGGGG CCAACCCAAA ATAAACATGC GAACTTATTA TTTATAGGCT	11100
CCATGCCACAC TTCTGAAAC CTCCATGGTC CTACTGGTTC CTGATTACCT CCACTCAATG	11160
AGAGGCAATT CATTACTGAA TGAGCCATAA GCGCCTCTTA TTTCGAGAGG GGGATGGCAG	11220
GACTCAGTCG AGGAGAAGGA CCCCACCCAG GCAGCCTGGG CCCCTCGGCT CCTGTACTTA	11280
TTTACTGCTG GGTACTTCCT AGCCCAGCAT GTAATTACTG GTTCGTTCAAG TCATTGTTT	11340
AGTAAATGTT TCTTGGGCAC CTACTACATA GGAGGCACAG GTCAAGGCAC TGGGGATATT	11400
CTTTCTACCC ACCCCCTCCCC TCCCTACACT GTGATTAGGG ACTGACCGAT C	11451

## (2) INFORMATION POUR LA SEQ ID NO: 3:

- (i) CARACTERISTIQUES DE LA SEQUENCE:
- (A) LONGUEUR: 1834 paires de bases
  - (B) TYPE: acide nucléique
  - (C) NOMBRE DE BRINS: double
  - (D) CONFIGURATION: linéaire
- (ii) TYPE DE MOLECULE: ADN (génomique)

## (xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 3:

ATTTTTTTTT TTTTTTTG AACGGAGTCT CACTCTGCCA CCCAGGCTGG AGTGCAATGG	60
CGCGATCTTG GCTCACTGCA ACCTCCGCCT CCCGGGTTCA AGTGATTCTT CTGCCTTAGC	120
CTCCTGAGTA GCTGAGACTA TAGGTGCCCG CCACCACGCC CAGCTAATT TTGTATTTT	180
ATTAGGACGG GGTTTCACCA TATTGCCAG GCTGGTCTCG AAATCCTGAC CTTGTGATCC	240
GCCCCACCTCG GCCTCCAAA GTGCTGGAT TACAGGTGTG AGCCATTGCG AGCAGCCCAG	300
AACTCAATT C TTAACCTTTA AAGTATGATG AGAAGAAGGA TCAAGCCCTC ACCAGCCCAT	360
TTAAGGAGTT TAGGCTCAGT CTTGAGGATG TGAGAAGTCA TTGCTATTGG GTTTCACACT	420
GAGGTTAAC A GGTGAAGTCA GCATTTGGT AGTTCACAGC AGCTGCAACT CTTTGTATTT	480
CTCTGATACC TCCTGTCCCA ACCTACATCA GGCCTCCCT TCTTCCTGCT TCCTTAATT C	540
CTCCATTTTC CCACCAGATG GAAGGACTGG AGCTTTGTGG ACAAAAGATGA GAAGGCCGT	600
CTGCAGCACC AGGTCACTGA GGATGGAGAG TTCTGGTGAG TCCAGAACCC AGGAAGACCC	660
AGAAGGGTAA GGGTGGGAA GACAGGGAA ATCTCAGACC TCAGTCCCCA GCTAAGGTTA	720
TCAGATTCCA GCCCTTGGGA GATCTTGGCT GTGTTCTCCT CCAGCCCAAG GCCCAGCAAG	780
GATGAGGTTTC TGAGAGGAGC CTTCCAGGCC ACAGGGACAA TGAGCCAGG ACCAGCCAA	840
CATGACATGG CTCTTGCTC CTGTGTGCC C TCCGCCACA CACTCTATT C CAGCCACAGG	900
CACCCCTGGCC TTAGCACAAT TCTTTCTGA GCCTAGGAAG CTCCACTTAC CCTGATCTTC	960
CAACGTCAAC CTCACCCCTCT CTCAGGTTGT TTCTATTCA GCTTCAAGTC TCAGCTTAAG	1020
GAGAATTTC AAGTCTCAGC TTAAGGAGAG CCCCTAAGT TCCCCGAGGA CTGGGATTAA	1080
TTTATGATGC TCATCACCC T AAAATTGTT TGCTTAAGCC GGGCGCGGTG GCTCACCCCT	1140
GTAATCCCAG CACTTGGGA GGCGGAGGTG AACGGATCAC GAGGTCAGGA GATCGAGAAC	1200

ATCTTGGCTA ACACGGTGAA ACCCTGTCTG TACTAAAAAT ACACAAAAAA AGTAGCCGGG	1260
CGTGGCAGCG TGCGCCTGTA GTCCTAGCTG CTGGGGAGGC TGAGGCAGGA GAATCACTTG	1320
AACCTGGGAG GCAGAGGT TA CAGTGAGCCC AGATTGCGCC ACTGCACTCC AGCCTGGCG	1380
ACAAGAGAGA CTCTGTCTTG GAAAAAAA AAAAAATGTG GTCTTAGTTT AATGTCAAGG	1440
GAAAGGTTTT GGGTGTTTTT ATTACTTTAT TTTTATTTA AAAACTATAA TAGAGACGGG	1500
CCTCGCTATA TTTCTGGGC TGGTCTCAAA CTCCTGGCT CAAGCGGTCC TCCCACCTTG	1560
GCCTCCAAA ATGCTGGCAT GTGGGCCTGG TCAACATATG GGACCCCAAC TCTACAAAAA	1620
ATTTAAAAT TAGCCAGATG TGGTGGCGTG TGCCTGTAGT CCCAGCTACT TGGGAGGCTG	1680
AAGCAGGGGG TCACTTGAGC CCAGGAGGTT GAGGCTGCAG TGAACATATGA TTGTCGTTCA	1740
CTTTTCTTCT GAACGTGAGA TTAAGTGTAG TCAGCAATT GGCTTAGGAT TATTTATTCA	1800
GAATTTTAA CCGTCACGTT GCGGCAAACC AGGT	1834

## (2) INFORMATION POUR LA SEQ ID NO: 4:

## (i) CARACTERISTIQUES DE LA SEQUENCE:

- (A) LONGUEUR: 14664 paires de bases
- (B) TYPE: acide nucléique
- (C) NOMBRE DE BRINS: double
- (D) CONFIGURATION: linéaire

## (ii) TYPE DE MOLECULE: ADN (génomique)

## (xi) DESCRIPTION DE LA SEQUENCE: SEQ ID NO: 4:

AGGAGGTGGA GGTTGCAGTG ACCCAAGATC ATGCCACTGC ACTCTAGCCT GGGCAACAGA	60
GCGAGACTCT GTCTCAAAAA ATACACACAC ACACACACAC ACACACACAC ACACACACAC	120
ACACACATAT ATATACACAC ATATATATAC ACACACATAT ACACACACAC ACGTCTGTAT	180
ATATATGTGT GTGTGTATAT ATACACACAC ACACACACAC ACACACACAC ATATATTCTT GTAGAGCTAT	240
GTGTGTCTCC TGTGCTATTG AGCATGAGCC CTTTTTTTTT TTTTTTTTTT TTGAGACAGA	300
GTCTCACTTT GTCGCCAGG CTGGCATAAC ATGGCGCAAT ATCGGCTCAC TGCAACCTCC	360
GCCTCCTGGG TTCAAGTGAT TCTCCTGCCT CAGCCTCCCA AGTAACTAGG ATTACAAGTG	420
CCCCGCATAA TGCTCAGCTA ATTTTTGTAT TTTCAGTAGA GATGGGGTTT CACCATGTTG	480
GCCAAGCTGG TCTCAAACTC CTAGCCTCAG GTGATCCACC TGCCCTAGCC TCCCAAAGTG	540
CTGGGATTAC AGGCATGAGC CACAGCACCC TGGTGAGCAC TAGAGCTTAT TTCTTCTATC	600
TAACTGTATT TTTGTATCCA TTAGCCACCC TCTTTTCATC CTCCCCTCTC CTTCCCTTCC	660
CAGCCTCTGG TAACCACTGT CTGCTCTCTA CTTCCATGAC ATATGCTTTG TTTTAGCTCT	720
CACATATGAG TGAGAGCATG CGACATTTAT CTTCTGGCC CTGGCACATT TTTGAATCAT	780
TGTTAGAAAA GATGATGGTT TGGAGTAGAT ACATCAGAAC TGACAGCGTT TGCCCTAAAA	840
AGGAAAGACA GGCTCCTCTG GGACCCCTGAC CAAGTTCTG TGAACATT TATTATTGTG	900
CTGTGTTAGT CCTGGGGTCT TCCGTTCCCA GCCCTCCTCA CCTGCTCCCA TATGGCTCTC	960
TCTCTTCTTC CAACCTCTCA GGATGTCCTA TGAGGATTTC ATCTACCATT TCACAAAGTT	1020
GGAGATCTGC AACCTCACCG CCGATGCTCT GCAGTCTGAC AACCTTCAGA CCTGGACAGT	1080
GTCTGTGAAC GAGGGCCGCT GGGTACGGGG TTGCTCTGCC CGAGGCTGCC GCAACTTCCC	1140
AGGTGGGAGA TGCTCTTGAT GGGGGGAGGG TCTAAGCCGA AAAAGTTCCA GGCAGAAGAA	1200

GCCTAAGTAG TGCTTATTAA GTCTCTCTGT TCCACACGTC CACTATCTTA TTAAACCTTC	1260
CCGTTTTAC TGAGAAGGAA ACCACCATGC TGAGAAGTTT GCAATAGGGA GCTGGGTAGC	1320
AACTTTGGAA GCAGGAACCTT GTGGGAACAA TGCAGATGCT GCTTGGACTT ACGATGAGGT	1380
TATGTCCAGA TAAGCCCATC CATCTTTGA AAATACCCCTA AGTAAAAAGT GCATCCAATA	1440
TGCCCTAACCC CCCAAACCTC ATAGCTTACC CTGGCCTACC CTCAAACATT GCTCGGAACC	1500
CTTGACCTTA AGCCTAAAGT TGGGCCAAAT CATCTAACTC CAAAGCCTAT TTTACAAAGA	1560
AAGTTGTTGT AATATCTCCA TGTAACCTAC TTAATACCTG TACCTAAAAA GTGAAAAACA	1620
AGAATGGTTG TACGGTACT CGAAATCCAG TTTCTACTGA ATGTGCATCT CTTTCACATT	1680
GTAAAGTTAA AAAATTGTAG CGGAACCATC CTAAGTCAGG GACTGTGAGT ACTGTGTCAG	1740
TAACAGTAAG GGCACATTG GAGAACCAAG TTAGCAGCTG CTGCAATAGT TCAAGTCAGA	1800
GATGATGAAA ACCTAGACCA AGTCAGTAGC ACCAGAGATG GAGGGGAGAC AGCAGATTAA	1860
GGGAGAGCAT ATTGGGTGAT GTAGGGAAGG AAGAAGAATG ATGTCAAGAT TCCCAGTTGG	1920
GGACCTGACA ACATTGCAAC ATAAGACACA CAAGAAGATC GGGTGGGTGG CTCATGCCATA	1980
TAATCCCAGC ACTTTGGGAG GCAGAGCCAG GAGGATCACT TGAGCCCAGG AGTTCAAGAC	2040
CAGCACAGGC AACATAGTGA CACCTCATCG TTACCCAAAA TAAAAAAAAA AATGAGGTGG	2100
GAGGATTGCT TGAGCTCGGG AGGTTGAGGC TACAATAAAC TGTGATCATG CCACTGCACT	2160
CCTGCCTGGG TGACAGAGTG AGACCCCTGCC TCAAAAAAAAAA AAGACACACA AGAGAAAAAT	2220
ATCAGCGTGT TGTTTGTGTT TGGTGGAGTT AATTGTGGGG TTCTAGGGAA AGGAATTAG	2280
CTTGGGACAT GGAAAGTTTG AGGTTCTGT AGAGTGTCCC AGTGAAGATT TGTAATAGAG	2340
CATCGGATGC GCATATTAGA TGGCACTTGG TGATATGATA AGAACTCAAA AAATATTGAA	2400
GGAATAAAGG AAAGAAGAGG CCAGACGTGG TGGCTTATGC CTGTAATCCC AGCACTTTGG	2460
GAGGCTGAGG CAGGCGGATC ACTTGTGGTC AGGAGTTCGA GACCAGCTTG GCTAACATGG	2520
TGAAAACCCA TCTCTACTAA AGATACAAAA ATTAACCGGG GATGATGGTG GGTGCCTGTA	2580
ATCCCAGCTA CTTGGGAGGC TCAGTCAGAA GAATCGCTTG AACCCAGGAG GCGGAGGCTG	2640
CAGTGAGCCG AGATCGGCC ACTGCACTCT AGCCTGGCA ACAGAGCCAG ACTCCGTCTC	2700
AAAAAAAAAA AAGTGAGAGA GATTGAGGCT GGGATATATG GCTCAGGCAT CATGCCGTG	2760
TAGGGGGCAG TTAAAAAGCA GAAGTAAGAA AGATTGCCTA GGGAGGCAGG AACGGTGAGG	2820

TGAGAGGAGA AGAGGCCAG GACCAGATT TAGTCACCAA CAGCGTTAA GGGGCAGGTA	2880
AGGAAAACAA AACCATCAGC AAAGACTGAG AATGAAAGCC CAGAGAGGAA GGAAAAGCCA	2940
CACATACAAT CAGTACAGCT CCATCTGAAT AAAGGTAGCG CCCCCCCCCC CCCAAATCAT	3000
TAGAGAAATG CCTGATTCGG TTTCTGTGG ATTTTCCTA AGAACCTAGA TGTGGGAAT	3060
AGAAATAAAAT GGTTCCCTCT GTCTCATCCC CTCCCTGCC TCTGAGAGGA AGCTGTGATT	3120
GCGTGCTCCC TTTCTGGGG TGCGAGATACT TTCTGGACCA ACCCTCAGTA CCGTCCGAAG	3180
CTCCTGGAGG AGGACGATGA CCCTGATGAC TCGGAGGTGA TTTGCAGCTT CCTGGTGGCC	3240
CTGATGCAGA AGAACCGGCC GAAGGACCGG AAGCTAGGGG CCAGTCTCTT CACCATTGCC	3300
TTCGCCATCT ACGAGGTGTG TAGTCCTGAT TGGCTCCAGC CCAGGAAACA TACTTTCCA	3360
GAGAGGACGC TTCCAGGGGC TTCTAGAGGG GCCCTCTGCT TCCTCAATAC CAGTGACCCA	3420
CAGAGCTCCT GGTATCAGGA CCACTTGTGT TTGTAACAAG CAAAAAATAC CAGGGGGGGC	3480
ATTAGAGAGG CAGTGGAGCG GGCCTGGCAG AACAGGTGCC TGGGGTCAG GCTTCCGCAT	3540
GCGGGCTGCA GTTGCTGGCA TTGCCTTCGG CAGGCTCCTC ATCCTCATTC ACATCTGAAG	3600
CATCTTCCTT TCTGTTCTT CTCAAGGTTC CCAAAGAGGT ATAGCAGCAG CAGCGGCCAG	3660
CAGTTGTGTG CAGCACTACC CAGGGGGGCC CGAGTCTGTC TGTGGCTCGT CGAGAAGCTT	3720
CCTGGTGGGG TTTGTGGCA GGACTTGTGA TAGGAGAGGG CCTTGCCTGT TGTTATTCC	3780
CACTTGCAGA GCAGGTTGCC TCAGGGCATT GCATGACCCA TGACTACCAC CCCCAGGATG	3840
TGCACTTTCT CCCTCGCACC AGACACTGCA CGTCACACAC ATGCCCTTGC ACACTCACCC	3900
TCCTCCACGC TTACAGCCAC ACACACAGTC ACACAGACGC GTTCTGAGGG TGGCTGCCCG	3960
CTTGGGATGG AGGAATCACT TCCCTCAGAA CCCAGCCAAG TCCTCTAGGC CTCCCTGGGG	4020
CTCCTTCCAG CCTGAGGGGC TTGGGAGCTG AGGACAGCTG TTCTGGTAAG TGTCCCTGAG	4080
TGTGGGATG ACACATTCC ATTCACTCTG AATCACAACA GAAAAGGGAA GAGGAATTGA	4140
GGTAGGGAGC CTATTTAACC CTTGGGAGTC GGGAAAGTAGG GAGGTTGAAA CTGTGACATG	4200
GGTGACCAAGG GAGTTGGAA GGGACCCCTG GAGGTGGCTG TGGCAGGACA GGACGTTCCCT	4260
CCCGAGGGGC TCATGTGCCCT TGGGCTCTCC CCATCTCTCA GATGCACGGG ACAAAGCAGC	4320
ACCTGCAGAA GGACTTCTTC CTGTACAACG CCTCCAAGGC CAGGAGCAAA ACCTACATCA	4380
ACATGCGGGA GGTGTCCAG CGCTTCCGCC TGCCTCCAG CGACTACGTC ATCGTGCCCT	4440

CCACCTACGA	GCCCCACCAG	GAGGGGAAT	TCATCCTCCG	GGTCTTCTCT	GAAAAGAGGA	4500
ACCTCTCTGA	GTGAGTGCTG	GCCCAGCTT	CCCACGTGTT	TCTAAAAGCT	CACATGGCCC	4560
ACTCCAGAGG	TTGAAGGCAT	GAGGCAGCTA	GACACGTCTC	CTCCAGGGTC	CTTCTGCTGC	4620
TCCTGAGCCA	CTGGCACAT	TACCCCCATT	CATTCAATTCA	TCCATTCTGT	GATATTATT	4680
GAGCACCTAC	TATGTTCCAG	GCACGTGCCT	AGGCACTAAG	GATAGAGTAG	TGAAGTAAAC	4740
AGAAAGAAAT	CCCTGCCCTTC	ATGGAGCTTA	ATATTCTAAC	ATGAGACAAT	AATGGATAGG	4800
AAAAACATAT	GTAGCATGTT	AGATTTGGAG	AGGTGATATG	GAGCAAAAT	AAACTAGGGA	4860
AGAGGGATAG	GAGGTGTTGG	GGATGCTTGA	AATTTAGGT	TAGCATGGCC	AGGAAAGCCA	4920
CATCCTGTCC	CTGGCACCA	CAGATGAGCT	CATAGCCCC	GCCACTCTGA	TCTCTGCTCCT	4980
TGGAAGATGC	ACCAGGTCCA	TGGGTAGGTG	GCTGGTCAT	GCCTTGGGG	GGCTCTGAGC	5040
AATACTAACCA	AGAACCTGCC	TGCCTGGGCT	TGGCTGTCGG	GGATGGTGCT	GACATGGGGC	5100
TGGTTCTGG	GGTTGGGTG	TTCCAGGGGT	TCTCTAGAGG	CTGGTTCTGG	CTTGGCTGCC	5160
AGGAAGCCGT	GCACCAGAGC	AAACCGTCCA	CGGGCCTCCT	GCTTGCTTCT	GGTGACACTG	5220
AGACCCCACA	TGTCTGTATT	CCTCACAGGG	AAGTTGAAAA	TACCATCTCC	GTGGATCGGC	5280
CAGTGGTGAG	TGGTTAGAT	CTTCTGTGCG	AAAAGTCCAG	AGGGTCCCCT	TCCCTGACCA	5340
TGCAGGGGAC	AGATGGTGCA	GGGGAGAATG	GGCACTGGCA	GAGGGAATGG	GAGTCTGGC	5400
TGTGCTGAGC	AGTCCCTCCT	TGGCACTGCA	AATCCTACTT	TGGCATGGCC	AGAAGTAATC	5460
GGCCTTAAGC	ACCGGGGGCC	ATTGAGGCAG	TTCAGGGCT	GGGAAATATG	GAAGAGGGTC	5520
CTGGAAAGGA	GAAGCAATT	GAACAATCGG	AGGGAACAAAG	GCCACAGGAA	GGGATGACAA	5580
GAGCCGCAGC	GAACACTGGA	TTCTGAGACT	GGATAACATT	GGATTTCACCA	CATAGAGAAA	5640
AGAAAGTAAG	CTGGTCCCGG	ACCTGGTGT	GACACTTGG	TCCTCCACTT	ACCAGGGGG	5700
TGACCTGGAC	AATTTCTGTA	ATCCCTCTCA	CTCAGTTCC	TACTCACTAA	AACGGGGATG	5760
ATAATGTGCC	TTGCAAGGCT	TTTGTGAGGC	TTCATCAATG	AGGTGATGTA	TGTGAAGTGT	5820
CTGGCACAGC	ATGGGCACTC	AAACAGAGGT	GCTTTTCAC	ACTTTACACC	TTACAAGGTA	5880
CTTTTCACAT	GTGTCATCGC	GATACTTGC	AGGTTGCTGA	GAGGTAGATG	GGGTTATAAT	5940
CCCTGGTGT	CAAGAAAGGA	AGCAGAGGCT	CAATGGGTT	GAATGACTTC	TCTGAGTTCA	6000
CAGAGCTCAG	TAAGTGGCAG	GGTTGGAAC	TCACATTCA	ACTCTCTGAC	TCCAGACTTA	6060

GGTTTTCCG CACCTCCACG CTGAGGCCAG CCCCAGGCAG TGAGAAGCCC AAAGTCCGAA	6120
GCACAGACTG CTGTGTGTTG GGCTCTGTGT GTTGAGGAGT CTTGTGACTG CCTTGGGGCT	6180
TTGGGCTGTA GTCAGCTGAC AGTCCTTGT GCTCTGTGGG GATGACGTAG GCCAATGGGA	6240
GGACAAATGC CCCTCTGAAC TGTCTTCTGG GCAGTGACAG TCATGGTCAT AATCCTGACC	6300
CTGAGCCAGT GCCAGGTCTC CAAGTGCCTT CTGAATGACC ACAGGCGATT GGTTTAGTG	6360
GTAGGTGCGT GGGGATCTGT TCTGGTCATC TGGATGCTGG TCATCGGGTG CAGTATTGAT	6420
CAGGACCTGC AAACCCAAAA GCTTATGGGA GCTGGCACGT CACGTGAGTA GAGCAGGCAG	6480
GTGCAGGGTT TTTGATGTCC CTGCACTGAC ACAGTTGTCT GCAGTTCTCC AATTTGACAT	6540
TTGGGCTCCA GTGTCGAGGG TCAAACAAGG AATTTGGGG CGTGGGCCAA ATCTGGGAAG	6600
ACACAGGGAG CAGGGCCCTT TGGCTCAAGC TGATAGTTGC CGCAGGGATT ACCAGGCCA	6660
GGGCAGCCTG CCACAAGCTG GGGCTTTAC CAAAGAAAAT CTCCCTATGT TAAATGCTTG	6720
CTCAAAAATT TTTAAAAAAAT ATTCTGTAAG TCAAAATCCA TTCTTAGGTC AGTTTGAGAG	6780
AGCCATGTTT TTGGTGTTTT AGTAACCAAT TTCATTTTTT TATTATTTAT TTATTTGTTT	6840
ATTTTGAGA CGGAGTTCA CTCTTGTAC CCAGGCTGGA GTGCAATGGC ATGATCTCAG	6900
CTCACTGCAA CCTCCGCCTC CCGGGTTCAA GCAATTCTCC TGCCCTCAGCC TCCTGAGTAG	6960
CTGAGATTAC AGGTGCCAC CATCACGCCCT GGATAATTTT TGTATTTTTT AGTCGAGATG	7020
GGGTTTCACC ATGTTGCCA GGATAGTCCT GAACTACTGA CCTCAGATAA TCCGCCACC	7080
TCAGCCTCCC AAAGTGCTGG GATTACAGGC ATGAGCCAGC ACCGCCCCGCC ACCAATTCA	7140
TTTTTAAAAA AAGGAAGAAA GAAAACCTTA GCCAGAAGAT CTTTTCTT CCCATATGCA	7200
GTAAGAGTAG ATTATAAAAA CAAAGTCAGA GCAGTCACTG GTGCTGGGC ATGGAGGAGA	7260
AAGAAGAATT CTCTTCTCCC TTCAACCTCC ATGCCCTTT TTGGCTCCAT GTGATTCA	7320
TTTCTGGACC CTGGAGCCCC ACCCAAGCT AAAGACCAGG ATACAGGGAA GCCACAACCA	7380
CTGGCGGTTC TGAGAACTTA CTTTCACIT ATTCTGCATT TACTGTTCC TTTCTTATG	7440
CAGAAAAAGA AAAAACCAA GGTAGGTGTG TGGGTAGAGA GCATGAAGTG TGTGTACTCA	7500
TGCATATGTA TGTGCATGCA TGTGAAGTGT GCATGTGTGA GCTCATATGC ATCCATGCAC	7560
CAGACTTGCC TCTTCCTCCC CCTCCTTCCT GAGCTCTGC TGGGGCCGAG CGTGCAGTAA	7620
TGACAACTAC GATTTGCTGG GGGAAAGGCTA CGTGCCAAAGC ACTCTTTAG GTGCTTCCA	7680

TGATTAATTCTTCACAACAGCCCTATGAGATTAGTA	CTATAACTATCCCCATTTTC	7740
AGAGGGAGAA AAGGTACAGA CTTGACTAAC TTGCCAAGG CCACACAGCC AGAGAGGGC		7800
AGAGCCAGTA CTTAGAGCCA GGCAGTCTGG GTCCAGAGTC CGTGTCTGA ACCACAAGAG		7860
GCCATCATAC GCCATCAGAT TTGGTGCTAG CATTCTGGT GGTGCCTGGT GGTGATGGAT		7920
CCATCACAGG GGTCCCTCAG GTACTGGTGC TGGCCCAGAC CAGAGCTGAC ACTCCTCAGG		7980
CACTACCACA TTCCAGGCAC TGTGCTTGGG GTCAGTCCT CTCTTTTTT TCCCCCCAA		8040
TTATAACAGT ATCTACAAAG TAGGTGCTGT TATTTTCCC CTTTCACAGG TGAGATAGAC		8100
TCAAAGAAGT GAACTTGCCTC AAGGAACAGA ACTAATGAGT GGGGAAAATG GAACTGGAAA		8160
CCATGTCTGT TTACTCCAAA ACCTGTGTTT CTTGCCCTCT TTCTCTGATG CCAGCCCCCT		8220
ACACTTCAAG GCCTGTGTTG TCCAGACCCA CACTGGGCC TGCCAGTGTG TGCCTGGCAG		8280
GGATGCTCCA TGGCCACACC ATATCCATCC TACACATCCC CCCTCAGACT GTGACCTCCA		8340
TTTGCTCTGG GATCCCCACA AGCTTCAGCT GCTTGAGCAA GACACTGCTT AGAAGGCAGA		8400
GCAAGCCAAG GCCTCTGGGG CCTGCTGGGA GCCAAAGCTG GGGAGCCGTT TCCACGGGTC		8460
TATCTGCTTG AGCTGTCTTA GATGAGCAGC ATGGAAGGGC AGTGGTGCAT GAGTCCAGGC		8520
GGGCTGCTTT TCTGCTCCGA GAGGCTCTGC CTGCCAGTT GTTCTCTGCA TTGCAGCCTC		8580
AATCCCCACA GCCTTGCCTT CCCCCGGCTT TCCCTACAGG TGACCCGCAT CCACAGTGT		8640
GGCACCATGC AGCAGCCGCT CTCCGTCTT TTCATATCCT TGTCACTTGC ACGAGCATGT		8700
CTTGAAAATA TCCCTTGTCTT GTGTAGCATC TTAAATGTTT TTGCAGTATG ATTTTGCATT		8760
CAGTATCTCA TTTGATCCCC ACAAGAGCCC TATGAGGAGG GAAAGCAGAT TTTACCATTA		8820
AAGGATGAGT AAACTGAGGC CAGAGAGGAT ATTTTGGTT TTTTTGAGA CAGTCTCACT		8880
CTGTCACCCA GCCTGGAGTG CAGTGGCTTG ATCTGGCTC ACTGCAAGCT CCACCTCCCA		8940
TGTTCACACC ATTTTCTGC CTCAGCCTCC CAAGTAGCTG GGACTACAGG CACCCACCAC		9000
CACACCCAGC TAATTTTTT GTATCTTAG TAGAGATGGG GTTCACCCA GTTAGCCAGG		9060
ATGGTCTTGA TCTCCTGACC TTGTGATCTG CCTGCTTCGG CCTCCTAAAG TGCTGGATT		9120
ACAGGGTGA ACCCCCCCTGC CCGGCCAGAG AGGATATTC TTAATGAGGG GCAGGGCTGG		9180
GATTCCAGCC CAGTGTCTG ATGGCTCACC CACTGACCAT TCCACTAATC CGTGTCTTT		9240
TTCAATCTAA ACTTTCAAGGG TTGTAGAGGT TCCTTGAGG TGCCTCAGTA CTTCCATGGT		9300

GATGTGGGTT CTGAGGGCCA AGAGCTCTGT TCTCATTAAT CAGAGAAGCT TGTGTTTTA	9360
AAAACACCAT GTTTACTGCA GGAAATTAA TTGGACAGTG TTTCCATCTG GAAAAAAA	9420
AGTCTACAAA ATACTTGACA ATCACTGCAC TAGATCATGC TGCTTTAGC ATTCTTAGCA	9480
TTTCACGTGC TGAGCTCTCA ATACTCTACC ATGAGGAGGG ATGGAGTGGG TATGAAAAGA	9540
TAAAGAACTG AAGTCACACG GCTTGTCACT GGCAGAGATA GAGCTTGAAC CGAGGTTGAA	9600
GAGCTCCCGC CTATTCCCTT CCTCTTCTCA CTGGATAAAG CTGCTCCAAG AGAGGTGCTG	9660
CCTCAGTGTG CCTGTTCAGA CTGTAATCCT CCCTTCCTTC CTGCCTCCTC CCTCCTCT	9720
CCAGCCCCATC ATCTTCGTTT CGGACAGAGC AAACAGCAAC AAGGAGCTGG GTGTGGACCA	9780
GGAGTCAGAG GAGGGCAAAG GCAAAACAAG CCCTGATAAG CAAAAGCACT CCCCACAGGT	9840
GTCTGGGCAT GTGGCATGGG TGGGGTGGCC AGCACCGTAC AGGGGCTTCC TATGCGCTTG	9900
GGATACACAG GGGCTGGAGG CTTCCCAGGA GTTTGTCTTG AACATCTGGA GGTTTGAATT	9960
TGTCCCAGT ACCTTTCTT TCAGCAAGTT CCCCTGAAAT TTGGGCTGCT GCTTGGGTGA	10020
ATATCCCAGG ATGGGGGTTTC CATTCTAGGA GTGGACTGGC AGGCTGAGCC TCCCATGGAG	10080
CTGATCCAGC CAGGATACAG AGAAGGGAG GCAAAGGCTG AGACAGAACC AGCTTGAGAG	10140
CGGAGGCGCA ACTCTTGCT CCTGGTGGCC TTGAGCATT CACAATAGGG GGATAAAGGA	10200
TAGGAGCAGA AAAGTGGGCC TGACTTCAGA AATGGGGTCC TCTAGAGCTC ACAGGGAGGGT	10260
GTTAGATTGG AGTGGGAGCT TAGTGGAGGT GAGCCTAGA GGCAGGAGTC TCCAGACCAA	10320
TCCAGGCCCC CTCTTCTATC CGGGGGCCCC TCTTCTATCC AGGGCCCTC TTCTGTCTGG	10380
GAGCCCCCTCT TCTATCTGGG GCCTCATGCA GTGGGGCTA GGGGAGGTTTC TCTGAGGACT	10440
TGGCCTTGAT GACAGGGTGG CTGGAGGAAT CAGAACGGTC AGACCTTCTT TGACCTGCGG	10500
GCACCTTTAG TTGGAATGCT CAGGCCCTGGG ATGGTGGAGG GGGCTCTTGC AGGTGGGGAC	10560
TGGGGTGGCG GGGAGGAGGC TGTATGGCCG CCATATCTCC TTTGGCTGGG GGCAGGAGTC	10620
CTGGAGAGGT GTGAAGAGTC CCTGAGGCCT CGATGCATCT CACTCCAGCT CACCAGGTCT	10680
GCATTTGCCCT GTCCCCAGCT CCTGCTGCCA CCCCCGGCCG TTTTAGGCAC TTGGCTCCCT	10740
TGGCCCCAGAG GAGCTTGCCT CACAGGCCTG TGCACCTCTG ACCCCTGTGA ACCAGTTTC	10800
CTTTGTGCCT CCACAGCCAC AGCCTGGCAA CTCTGATCAG GAAAGTGAGG AACAGCAACA	10860
ATTCCGGAAC ATTTCAAGC AGATAGCAGG AGATGTGAGT ACCTCCAAGC CCAGGACGCC	10920

CACAGGTGCT	TCCTTCTCTC	CTGGATTAAC	TGCTCAGATT	ACCAATTATT	TCATTATTGT	10980
TTGGTAGAGG	TCACTTTGGA	CTTCGGTGGA	GCCAGGGGAT	GTGTGCCTAG	CACACAAATC	11040
CACAAGCCCT	TGAGTTTG	ACTGCCACGT	CTGCTGGGGG	GCTCAGAGGC	CTTTTGCTC	11100
TGAGCTGCC	ACGGTGGTCC	TGATAGCTGA	GGTGCAGTAT	CTGGCCCCCT	GTCTTCCTCA	11160
GAAAAGCCCC	AGCTTCCCCT	GACATAATAG	CACCGACAGG	GATTTACAA	ACACAGCCAG	11220
GTGGAATTG	TTTGCAAAG	TGTCCGCGCC	AGGAGCTGCT	GTACTCCTGA	ACCATGACCC	11280
TCCTCTCCCT	TCCTCCTCAG	GACATGGAGA	TCTGTGCAGA	TGAGCTCAAG	AAGGTCCCTTA	11340
ACACAGTCGT	GAACAAACGT	GAGTTGCTCA	AACCAAATGG	GGGTGGGGTG	GGTGGGGACT	11400
CCCGTTGTCT	CAAAGCAGCT	CCTCACTCTT	CTCCATCCCC	CCAGACAAAGG	ACCTGAAGAC	11460
ACACGGGTTC	ACACTGGAGT	CCTGCCGTAG	CATGATTGCG	CTCATGGATG	TATCCTTCCT	11520
GCCGCCCTT	CCCGACCCTC	TGTATCAGC	CCACGGGGGC	CAAGGCAACA	TACAGGGTGC	11580
CCAGTCAGGC	AAAGGGCCCT	AATTGTGCC	CAGGGAAACT	TAAGGAGACC	CTGATTCAAGA	11640
ACATCTTGGA	TACTCGTCTG	AAAGGGTTG	TTAGAGGCGG	AAGGGAGGA	TGTTGGGTTG	11700
TAACTGCCCT	AACCCCTGTG	CTTCTCTCAG	GCCTGGGATC	CTGCCAAGC	AAAAGTGGTC	11760
CTTAGGAGAG	CGGCTCCTGG	GTTACAGAGT	AGGCGCAATC	TCTGACTGGT	GGTGGAGTGG	11820
AGGGGAGGGT	TAAATAGTAC	AACAGGGCAG	TGGTAGGAC	AGCCCGGAGT	CTCCTAGACC	11880
CTCCCTCCAA	ATCCAGGGGG	ATTTGCTGT	GTGCTGTGA	GCCCTGACCT	CCCTCCTCCA	11940
GACAGATGGC	TCTGGAAAGC	TCAACCTGCA	GGAGTTCCAC	CACCTCTGGA	ACAAGATTAA	12000
GGCCTGGCAG	GTGGGAAGAG	AAAATGAAGC	GTGGGAGTCA	AGAATGGGGT	TGATTTGGAG	12060
ATTCAGTGTG	TGACCTCCAT	CCTCAAATT	TCTATTGCCA	GAAAATTTTC	AAACACTATG	12120
ACACAGACCA	GTCCGGCACC	ATCAACAGCT	ACGAGATGCC	AAATGCAGTC	AACGACCGAG	12180
GTGCTGAGAA	GGAAGGGGTG	TCAGGGATGT	GGACCCGAGA	CGGTGGGAGC	AGGAATGGGA	12240
GGGGACTAGC	TACTAGGGCC	CCACTAGAGA	AGGAGAGGGG	AAGGGCTTCT	CACTTTCCCT	12300
TCCCAGGTCA	CAGAGTGTCC	GAGAGGCAGG	GAAAATAGAA	GACAGGCCA	AGGCCTCCAG	12360
CTCCACGTCC	ACCTCTAAC	TGGTCCCTC	CACAGGATTC	CACCTCAACA	ACCAGCTCTA	12420
TGACATCATT	ACCATGCGGT	ACGCAGACAA	ACACATGAAC	ATCGACTTTG	ACAGTTTCAT	12480
CTGCTGCTTC	GTTAGGCTGG	AGGGCATGTT	CACTAACTGG	GAGAGGGGGG	CTGCCCTCTG	12540

CTCTCTTGCA	GGGGCAGTTG	TGGCAACAGG	CATCTCACCT	GATAATCTCC	AGTCTGCTCC	12600
ATCCAGGCTG	AACAAGGGCC	AATGACCTCT	TTAGGCCAG	AATGGGATGG	CAAAGGGAGG	12660
GTTACTGGTG	ATTCTCTGCC	TGCACATCTT	TGTGCTGATG	AGGGACAGCA	CTGGGCACAC	12720
GGTCCTCTGA	GGGGAAGTTA	CACTAGTAGA	GGCGGAGTGC	GCCTGTAAC	GGCCTCTGGC	12780
CTGTGCATTC	TTTCACAGGA	GCTTCTCATG	CATTTGACAA	GGATGGAGAT	GGTATCATCA	12840
AGCTCAACGT	TCTGGAGGTA	AAGCATAGGC	ACAGCACATT	CCCCCTACAC	ATTAAAAC	12900
AAGGTGGAGG	GGTCAACGGG	CGGGACTGGA	CCCAGGGTGT	GCTCCTCATT	TCCACACAGT	12960
GGTGGAGGGA	AGGGATAGGA	ACAGAACATG	GAGGGAGGCT	CAGCAGGCTC	CCAGGACACA	13020
TGCACTTGAG	GCCCCAAAGG	ACCTCTGCTC	CCCCAGTCAC	TTGATGCGGG	AAAACATGCA	13080
CCTTCTTAGG	GAAGATCTAG	GAGAAAGGAA	ACAGTAAGCC	ACTGCTTCTT	GGAAAATCTT	13140
CTGGGGGTCT	GACCTGCTGG	GACTGTTCCC	TTTCCTCTTG	CCCCGTAAGA	TTCTAGGGC	13200
GGGGGGGGGG	GGGGGTCACT	CTTTTCTGAT	CTACATTCTG	ATCTTGGAC	TTCTTCAGT	13260
GGCTGCAGCT	CACCATGTAT	GCCTGAACCA	GGCTGGCCTC	ATCCAAAGCC	ATGCAGGATC	13320
ACTCAGGATT	TCAGTTTCAC	CCTCTATTT	CAAAGCCATT	TACCTCAAAG	GACCCAGCAG	13380
CTACACCCCT	ACAGGCTTCC	AGGCACCTCA	TCAGTCATGT	TCCTCCTCCA	TTTACCCCC	13440
TACCCATCCT	TGATGGTCA	TGCCTAGCCT	GACCCTTTAG	TAAAGCAATG	AGGTAGGAAG	13500
AACAAACCCCT	TGTCCCTTTG	CCATGTGGAG	GAAAGTGCCT	GCCTCTGGTC	CGAGCCGCCT	13560
CGGTTCTGAA	GCGAGTGCTC	CTGCTTACCT	TGCTCTAGGC	TGTCTGCAGA	AGCACCTGCC	13620
GGTGGCACTC	AGCACCTCCT	TGTGCTAGAG	CCCTCCATCA	CCTTCACGCT	GTCCCACCAT	13680
GGGCCAGGAA	CCAAACCAGC	ACTGGTTCT	ACTGCTGTGG	GGTAAACTAA	CTCAGTGGAA	13740
TAGGGCTGGT	TACTTTGGC	TGTCCAACTC	ATAAGTTGG	CTGCATTTG	AAAAAAGCTG	13800
ATCTAAATAA	AGGCATGTGT	ATGGCTGGTC	CCCTTGTGTT	TTGTTGTCTC	ACATTAGAT	13860
ATCAGCCATG	CATGACTGAA	TGGCTTCAA	TCATATACTC	ACCTATCACC	TACAAGAGAA	13920
CAATGAAAAAA	CACACACAAA	AAACAAATCT	TGAATTGT	AATCATGCCT	ATTGCTATTT	13980
CTTGAGCATA	AGAATGGCTC	AGATACTTC	CAAGACATAA	AAGGAAGGCA	GAGGAATAGT	14040
TGTTGCTGTA	AAAGACATCA	AGAATAATG	GGGTCACTGTA	CAACGGGAGG	GGCCGGTTAC	14100
CTGAATAATG	GAGTGGAGAT	TGAGCTATCC	TAGCTCCTCT	GCTCACTAAC	TGACCTGTCC	14160

CATGACCGTG GACAAAACCC TGAACGCAGC TGTTTGTTC CTAAACTTCT CTGGACCATG	14220
GCCTGCGGCA TATCTATAAGG CATCCTGTGT TTTCCACCCA GTTTCCTTCT TCCTCGCTAA	14280
GCCAACGTGG AAAGGGCTGG CCGTGAATAT GCAGACAAGG TAACGAAAGT AAACCGTCAA	14340
TTAGTAAAAG TACTTCATTT TCCTCTTGT A TTTGCTTCAT TCTTGCTTCA CAAAGTTACG	14400
AAGTCCACAG CTTTATACCA AAATGTAAGA AGGCTATTTG CTTATAAAACA TTTTGAGTCA	14460
GGTGTCAATCT GATTTCATTC TTCTAATCCA TATTCAATAT TAAAAAAATCA GAAACCAAGG	14520
GTGCTGGAGC AGCTCTAGGG CATATATTTC TCTTAAATAG GAGAAAGATT TTCAACAGCT	14580
TTTCCTCCTT GACCCCCCTCC TTTCCAATT TATTTGGGTC ACTACCTTGA ATTTAGAGTG	14640
AATCTGGGAA ATGTAGTCAC CAGG	14664

